Univerzita Karlova v Praze
Přírodovědecká fakulta

Studijní program: Speciální chemicko-biologické obory
Studijní obor: Molekulární biologie a biochemie organismů

Jana Skopalíková

Využití metod next-generation sekvenování pro rekonstrukci fylogeneze polyploidních rostlin
Application of next-generation sequencing for phylogenetic reconstruction of polyploid plants

Bakalářská práce

Školitel: Mgr. Tomáš Fér, Ph.D.

Praha, 2015

Prohlášení

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 12. 8. 2015

Podpis

Poděkování

Na tomto místě bych chtěla poděkovat svému školiteli Mgr. Tomáši Férovi, Ph.D. za pomoc a cenné rady při psaní této bakalářské práce. Dále pak Mgr. Elišce Záveské, Ph.D. za připomínky a Mgr. Evě Skopalíkové, Mgr. Aleně Vašákové a Bc. Václavu Kubaljakovi za pomoc s odhalením překlepů a přehlédnutých chyb.

Abstrakt

Tato bakalářská práce shrnuje dostupné informace o aktuálně používaných metodách next-generation sekvenování (NGS), které v posledních letech zažilo velký rozmach. Výhodou nových metod je zisk velkého množství dat za mnohonásobně nižší cenu za osekvenovanou bázi oproti Sangerově metodě, avšak analýza těchto dat skýtá různá úskalí.

I když je v dnešní době již možné sekvenovat celé genomy jednotlivých organismů, pro fylogenetiku, která je postavena na studiu mnoha jedinců, zůstává tento přístup značně náročný. Proto v posledních letech vzniklo množství přístupů, jak účinně redukovat genom, např. sekvenování transkriptomu (RNA-Seq), cílené obohacení (target enrichment), restrikční metody (RAD-Seq, RLL, GBS), mělké sekvenování (genome skimming) a další. Každá z těchto metod má však několik výhod i nevýhod, které ovlivňují jejich využitelnost ve fylogenetických analýzách.

Dále se práce zabývá polyploidní speciací a specifiky studia fylogeneze u polyploidních rostlin – výběrem vhodných markerů, následným zpracováním dat a fylogenetickými analýzami. Poslední část je pak věnovaná mé budoucí práci na polyploidním rodu *Curcuma* L.

Klíčová slova

454 pyrosekvenování, Illumina, PacBio, RAD-Seq, Hyb-Seq, RNA-Seq, polyploidie, hybridizace

Abstract

This bachelor thesis summarizes available information about currently used next-generation sequencing (NGS) methods where a big progress was achieved during last few years. Great advantage of NGS is the ability to gain huge amount of data at much lower cost per base compared to the Sanger sequencing. However, there are various pitfalls in data analysis.

Nowadays it is possible to sequence the entire genomes of individuals. Nevertheless, this approach remains challenging when studying many individuals, e.g. in phylogenetics. Recently, several approaches for effective reduction of genome complexity arose: transcriptome sequencing (RNA-Seq), target enrichment, restriction digest-based methods (RAD-Seq, RLL, GBS), genome skimming (shallow sequencing), etc. Each method has both advantages and disadvantages that affect its utility in phylogenetics.

Furthermore, the thesis deals with polyploid speciation and particularity of phylogenetics in polyploid plants – selection of suitable markers followed by data processing and phylogenetic analyzes. The last part of the thesis is devoted to my future research of polyploid genus Curcuma L.

Key words

454 pyrosequencing, Illumina, PacBio, RAD-Seq, Hyb-Seq, RNA-Seq, polyploidy, hybridization

Obsah

Seznam zkratek

AFLP	amplified fragment length polymorphism
bp	pár bází (z angl. „base pair“)
cDNA	komplementární DNA
cpDNA	chloroplastová DNA
ddNTP	dideoxyribonukleotidtrifosfát
DNA	deoxyribonukleová kyselina
dNTP	deoxyribonukleotidtrifosfát
GBS	genotyping by sequencing
ITS	vnitřní přepisovaný mezerník (z angl. „internal transcribed spacer“)
MIP	molecular inversion probe
NGS	next-generation sekvenování
PCR	polymerázová řetězová reakce
PEC	primer extension capture
RAD	restriction-site associated sekvenování
RAPD	random amplified polymorphic DNA
rDNA	ribozomální DNA
RNA	ribonukleová kyselina
RRL	reduced representation library
SNP	jednonukleotidový polymorfismus (z angl. „single nucleotide polymorphism“)
SSR	krátké tandemové repetice (z angl. „simple sequence repeat“)
UCE	ultrakonzervované elementy

1 Úvod

Sekvenování DNA patří k důležitým molekulárně biologickým metodám, které se používají ke studiu dědičné informace různých organismů nebo skupin na různých taxonomických úrovních. Ke klasickým metodám, jako je Sangerovo sekvenování (Sanger et al., 1977), se počátkem 21. století přidává několik nových metod tzv. sekvenování nové generace (next-generation sequencing, NGS), které oproti původním metodám sekvenace DNA umožňuje paralelní sekvenování mnoha milionů molekul DNA, což významně zvyšuje množství osekvenovaných bází a snižuje cenu za bázi (Glenn, 2014). S příchodem těchto NGS technik, tzv. sekvenačních platforem (např. 454 pyrosekvenování, Illumina a další) těsně souvisí i vývoj mnoha nových molekulárních metod, které umožňují výběr různých úseků genomu, jež budou následně osekvenovány. Soubor všech (vybraných) úseků, které se sekvenují pomocí NGS platforem se obecně označuje jako DNA knihovna. V kontrastu se Sangerovým sekvenováním, kde bylo možno v jedné reakci sekvenovat pouze jediný fragment DNA (naamplifikovaný pomocí polymerázové řetězové reakce – PCR), umožňují NGS metody připravit knihovny zahrnující desítky, stovky, tisíce, ale i více fragmentů z jednoho nebo více jedinců, které lze následně sekvenovat na vybrané NGS platformě v jediné reakci. Tím se otevírá možnost pro rutinní velkoobjemové sekvenování organismů, včetně nemodelových, pro řešení nejrůznějších biologických otázek.

Své místo si NGS metody a next-generation sekvenování nachází i ve fylogenetických studiích, které se běžně zabývají větším množstvím jedinců a/nebo druhů. Využitím NGS metod dochází ke zvýšení vyšetřovaných lokusů z klasických jednotek až desítek (např. Fortune et al., 2008; Rousseau-Gueutin et al., 2009) spíše na stovky až tisíce (např. Cannon et al., 2015; Tennessen et al., 2014). Přestože s pomocí NGS platforem lze sekvenovat celé genomy, tato alternativa není pro fylogenetické studie příliš vhodná (stále vysoká cena, velká komplexita získaných dat), a proto se častěji využívá různých NGS metod umožňujících výběr určité části genomu (viz podkapitola 2.1).

Výběr cílové (sekvenované) genomové frakce se odvíjí od otázek, které si daná studie klade, množství jedniců, které zpracovává, míry vzájemné příbuznosti studovaných jedniců, ale i finanční možnosti studie/projektu. V současnosti nejpoužívanější NGS metody pro výběr sekvenované genomové frakce jsou různé restrikční metody (Baird et al., 2008; Van Tassell et al., 2008), tzv. cílené obohacení (Okou et al., 2007; Gnirke et al., 2009), sekvenování transkriptomu (Marioni et al., 2008; Morin et al., 2008), tzv. mělké sekvenování (Straub et al., 2012) nebo amplikonové sekvenování (Meyer et al., 2008; Bybee et al., 2011). Každá metoda má několik výhod i nevýhod

a je vhodná pro různé typy studií, proto je volba metody pro přípravu DNA knihovny pro následné sekvenovaní pomocí NGS klíčová.

Cílem této bakalářské práce je shrnutí poznatků o aktuálně používaných NGS platformách a NGS metodách využívaných pro přípravu DNA knihovny se zaměřením na ty, které se používají při konstrukci fylogenetických hypotéz a zhodnocení výhod a nevýhod těchto metod. Dále práce shrnuje současné poznatky o polyploidní speciaci, problematice studia fylogeneze polyploidních rostlin a hodnotí přínos a úskalí používání metod next-generation sekvenování oproti klasickým molekulárním technikám. Pro účely budoucí diplomové práce je na závěr shrnut i dosavadní výzkum u polyploidního rodu *Curcuma* L., pro který by bylo využití NGS metod pro rekonstrukci komplexních fylogenetických vztahů velmi přínosné.

2 Sekvenování DNA – klasické vs. moderní metody

Sekvenováním se rozumí proces stanovení pořadí nukleotidů (bází) v molekule DNA (deoxyribonukleové kyseliny). Od jeho objevu uplynulo již několik desetiletí. Mezi nejdůležitější z prvních metod patří Maxam-Gilbertovo a Sangerovo sekvenování. Maxam-Gilbertova metoda, též zvaná jako chemická, je založena na štěpení DNA molekul činidly, které rozruší chemickou vazbu v místě specifické báze. Radioaktivně značené fragmenty DNA jsou odseparovány podle své velikosti na polyakrylamidovém gelu, ze kterého je možné následně odečíst sekvenci bází (Maxam and Gilbert, 1977). Oproti tomu Sangerovo sekvenování využívá enzymu polymerázy, která syntetizuje komplemetární řetězec DNA podle templátu. Kromě 2´-deoxyribonukleosid-5´-trifosfátů (dNTP) obsahuje daná reakční směs také malé množství 2´,3´-dideoxyribonukleosid-5´-trifosfátů (ddNTP), které nemají 3´OH skupinu a způsobují terminaci nově vznikajícího řetězce. Od stejného způsobu detekce jako u předešlé metody (Sanger et al., 1977) se od roku 1986 přechází k detekci fluorescenčně značených molekul a od roku 1996 ke kapilární elektroforéze (Hutchison, 2007).

V roce 2005 byla na trh uvedena první next-generation sekvenovací technologie od firmy 454. Vždy s ročním zpožděním se k ní připojily další dvě metody – Illumina (původně Solexa) a SOLiD a do dnešních dnů ještě několik dalších, včetně metod sekvenujících jednotlivé molekuly, jako je PacBio (van Dijk et al., 2014). Oproti původním metodám se liší hlavně svou cenou za osekvenovanou bázi, která je o několik řádů nižší, a vysokým výkonem – jsou založeny na paralelním sekvenování mnoha milionů molekul v jednom sekvenačním běhu (Glenn, 2011). Při Sangerově sevenování vzniká mnoho různě dlouhých fragmentů jednoho úseku DNA (naamplifikovaného pomocí PCR), které jsou následně separovány podle své délky (Sanger et al., 1977). U metod next-generation sekvenonání se identifikují jednotlivé nukleotidy již v průběhu syntézy nového řetězce.

Vlastnímu sekvenování na některé z NGS platforem vždy předchází příprava fragmentů, které mají být sekvenovány některou z NGS metod. Tato fáze se obecně nazývá přípravou DNA knihovny a metody, které se k tomu využívají, jsou popsány v podkapitole 2.1. Následuje aplifikace fragmentů DNA knihovny specifická pro danou sekvenační platformu a vlastní sekvenování, jehož fáze jsou též specifické pro každou platformu. U většiny metod je identifikována sekvence nukleotidů při jejich začleňování do nově vznikajícího řetězce DNA, např. 454 pyrosekvenování (Margulies et al., 2005) a Ion Torrent (www.lifetechnologies.com[1]) detekují vedlejší produkty vzniklé při štěpení dNTP, zatímco Illumina snímá přímo jednotlivé fluorescenčně značené nukleotidy. Metoda SOLiD není založená na polymerační, ale ligační reakci (Valouev et al., 2008) a u PacBio nedochází k amplifikaci molekul templátu, protože tato metoda umí detekovat jediný

začleněný nukleotid (Eid et al., 2009). Níže jsou specifika jednotlivých (nejrozšířenějších) metod popsána podrobněji.

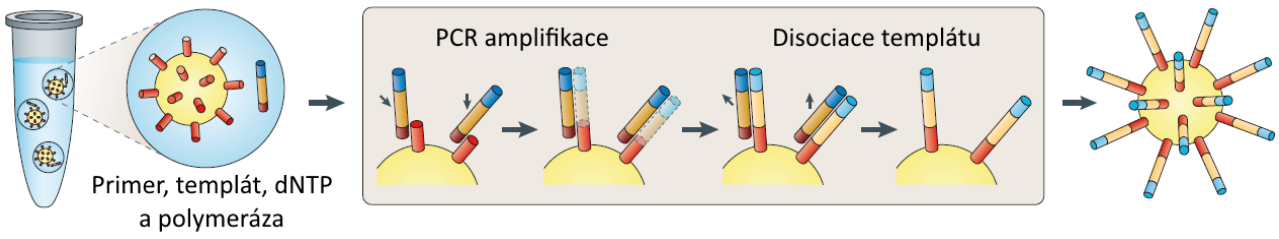
2.1 Aktuálně používané platformy next-generation sekvenování

Metody next-generation sekvenování můžeme rozdělit do dvou skupin – na tzv. druhou a třetí generaci. Do druhé generace se řadí ty, které vyžadují amplifikaci molekul templátu před vlastním sekvenováním, oproti tomu třetí generace sekvenuje přímo jednotlivé molekuly DNA (Glenn, 2011). Ze zde popsaných metod patří k druhé generaci 454 pyrosekvenování, Illumina, SOLiD a Ion Torrent a ke třetí generaci PacBio. K dalším metodám patří Polonator a Helicos, které jsou málo využívané, proto do této práce nebyly zařazeny. V současnosti se vyvíjí také několik nových metod, např. Starlight nebo nanopórové sekvenování (shrnuto např. v Egan et al., 2012).

2.1.1 454 pyrosekvenování (Roche)

První komerčně dostupnou metodou NGS se stalo 454 pyrosekvenování dnes patřící pod firmu Roche (Glenn, 2011), které je založeno na detekci pyrofosfátu odštěpovaného při sekvenování během syntézy milionů kopií templátové molekuly DNA uchycených na mikro kuličkách a namnožených pomocí emulzní PCR (Margulies et al., 2005).

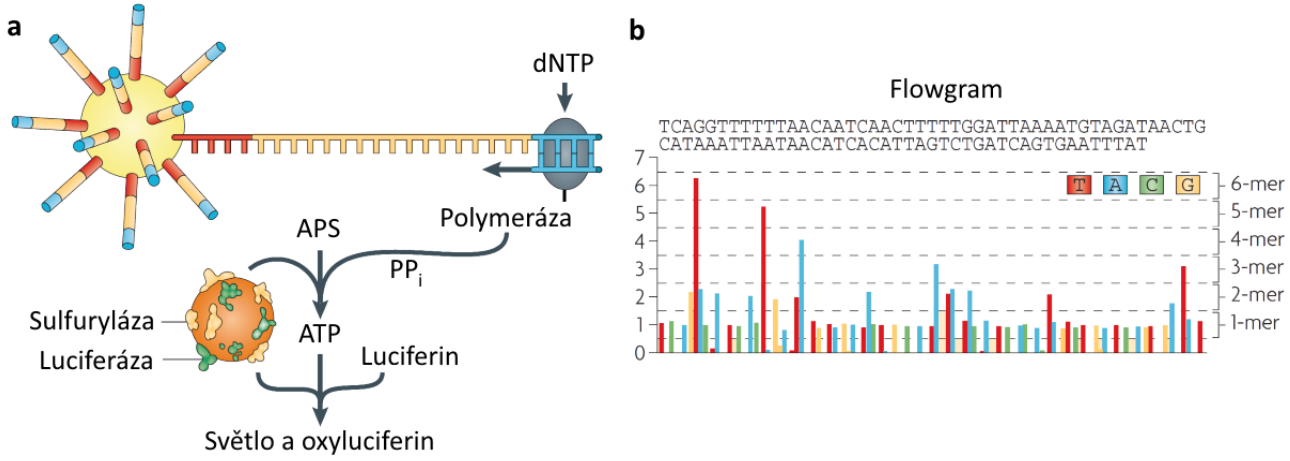
V rámci přípravy DNA knihovny jsou vybrány fragmenty určené k sekvenování a na oba konce těchto fragmentů jsou naligovány adaptory kompatibilní s primery navázanými na mikro kuličkách (tzv. „DNA Capture Beads“). Vlastnímu sekvenování předchází amplifikace DNA knihovny metodou tzv. emulzní PCR (www.454.com). Smícháním templátu s mikro kuličkami dojde k uchycení jednoho fragmentu na každou kuličku a po emulgaci s olejem se vytvoří vodné mikroreaktory, ve kterých PCR probíhá (Obrázek 1; Dressman et al., 2003). Vzniknou tak miliony kopií daného fragmentu navázaného na kuličce. Vzorek je nanesen na destičku (tzv. „PicoTiter Plate“) zkonstruovanou tak, aby se do jedné jamky vešla právě jedna kulička (www.454.com).



Obrázek 1: Emulzní PCR – emulze vodných mikro kapek v oleji slouží jako mikroreaktory pro PCR amplifikaci fragmentů, které asociují s primery na mikro kuličkách (převzato a upraveno z Metzker, 2010).

Na této destičce probíhá vlastní sekvenování. K sekvenování dochází paralelně ve všech jamkách, skrz které proudí jednotlivé dNTP v předem definovaném pořadí. Mezi jednotlivými

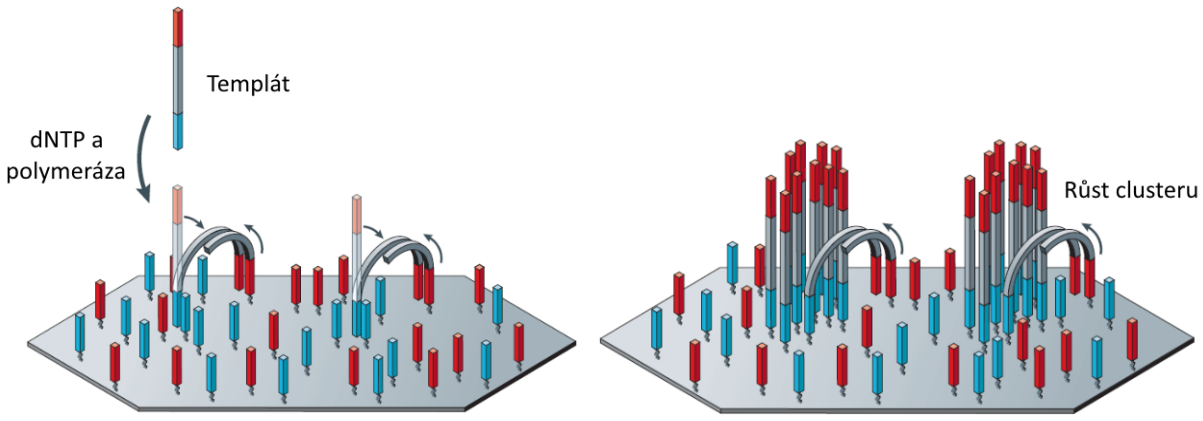
nukleotidy jsou jamky promyty apyrázou štěpící neodmyté dNTP z předešlého cyklu (Margulies et al., 2005). Při zařazení komplementárního nukleotidu polymerázou do vznikajícího řetězce dojde k odštěpení pyrofosfátu a sulfuryláza s luciferázou vytvoří pomocí chemických reakcí světelný záblesk (Ronaghi et al., 1996), který je zaznamenáván kamerou. Počet inkorporovaných nukleotidů je odečten z intenzity záblesku na tzv. flowgramu (Obrázek 2). Nejčastějšími chybami této metody jsou inzerce a delece u polynukleotidů, neboť v případě sekvence s větším počtem nukleotidů stejného typu se rozdíly v intenzitě záblesků z dané jamky stírají (Margulies et al., 2005).



Obrázek 2: 454 pyrosekvenování – a) po začlenění nukleotidu polymerázou do nově vznikajícího řetězce je z dNTP odštěpen pyrofosfát, který pomocí sulfurylázy a luciferázy dá vzniknout světelnému záblesku, jehož intenzitu detekuje kamera; b) vzniká flowgram, ze kterého je odečteno pořadí bází (převzato a upraveno z Metzker, 2010).

2.1.2 Illumina (Illumina)

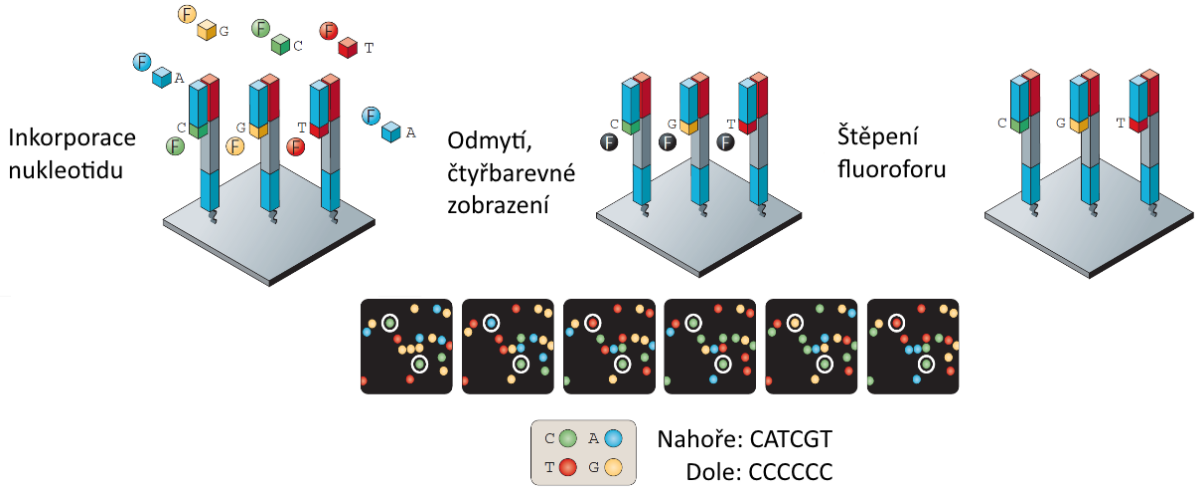
Solexa vyvinula druhou komerčně dostupnou NGS platformu, dnes známou jako Illumina, podle firmy, která ji vlastní nyní (Glenn, 2011). Tato metoda je založena na namnožení templátu pomocí bridge PCR na destičce (tzv. „Flow Cell“) a sekvenování během syntézy vzniklých clusterů molekul pomocí fluorescenčně značených reverzibilních terminátorů (www.illumina.com).



Obrázek 3: Bridge PCR – amplifikace na destičce pokryté primery, mezi kterými ohnutím fragmentů vznikají můstky. Několik cyklů PCR vytvoří clustery mnoha kopií templátu (převzato a upraveno z Metzker, 2010).

Pomocí adaptorů (naligovaných během přípravy DNA knihovny) se uchytí jednořetězcové fragmenty DNA na destičku pokrytou primery komplementárními k adaptorům. Dojde k ohnutí templátu, vytvoření můstků mezi primery a syntéze komplementárního řetězce. Několik cyklů PCR tak vytvoří clustery (Obrázek 3; Adessi et al., 2000) obsahujících okolo 1000 kopií původní molekuly DNA (www.illumina.com).

Vlastní sekvenování pak probíhá pomocí polymerázy a čtyř reverzibilních terminátorů a je detekován vždy celý cluster vzniklý v předchozím kroku. Po začlenění prvního nukleotidu a odmytí nezreagovaných dNTP je kamerou detekována příslušná fluorescenční barva, následuje odštěpení fluoroforu a 3´ skupiny způsobující terminaci syntézy a může být inkorporován další nukleotid (Obrázek 4; Bentley et al., 2008). Metoda téměř eliminuje chyby spojené s homopolymerními úseky (www.illumina.com). Inzercím je zabráněno reverzibilní terminací – možností inkorporovat pouze jediný nukleotid a delece jsou minimalizovány přidáváním všech nukleotidů najednou místo postupně (Bentley et al., 2008).



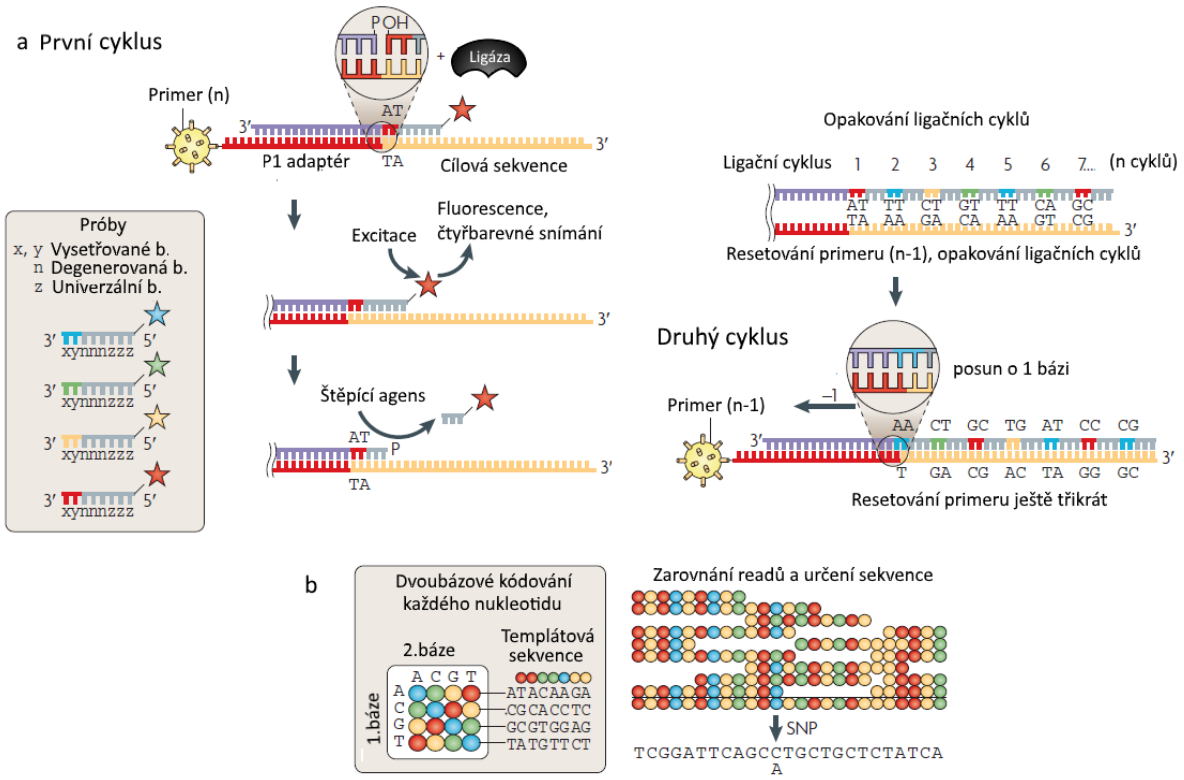
Obrázek 4: Sekvenování Illumina – inkorporování fluorescenčně značeného nukleotidu, odmytí nezreagovaných dNTP a detekce fluoroforu následovaná štěpením fluorescenční a terminační skupiny; ve spodní části je vyobrazeno čtyřbarevné snímání (převzato a upraveno z Metzker, 2010).

2.1.3 SOLiD (Life Technologies)

SOLiD, dnes patřící pod Life Technologies, se stalo třetí dostupnou NGS platformou (Glenn, 2011). Oproti ostatním metodám není založena na syntéze komplementárního vlákna DNA polymerázou, ale na ligaci fluorescenčně značených oligonukleotidů k primeru. Próby sestávají ze dvou specifických nukleotidů, tří degenerovaných a tří univerzálních bází a jednoho ze čtyř fluoroforů (www.lifetechnologies.com[2]).

Před vlastním sekvenováním dojde k naamplifikování fragmentů templátové DNA pomocí emulzní PCR (Dressman et al., 2003). Kuličky se zmnoženým produktem jsou poté kovalentně připojeny na skleněnou destičku (Valouev et al., 2008). Přidané próby hybridizují s templátem

a pomocí DNA ligázy jsou připojeny k primeru. Po odmytí ostatních prób se detekuje barva fluorescence a z oligonukleotidu se odštěpí poslední tři báze s fluorescenční značkou. Následují další kola hybridizace, ligace a štěpení (www.lifetechnologies.com[2]). Po té dojde k odstranění původního primeru a přiligovaných prób a templát hybridizuje s novým primerem (Valouev et al., 2008). Sekvenování se opakuje celkem pětkrát vždy s primerem o jeden nukleotid kratším, než byl předešlý. Každá báze se detekuje dvakrát a je dekódována podle barevného schématu (Obrázek 5). Díky tomu je dosažena přesnost sekvenování až 99,99 % (www.lifetechnologies.com[2]).



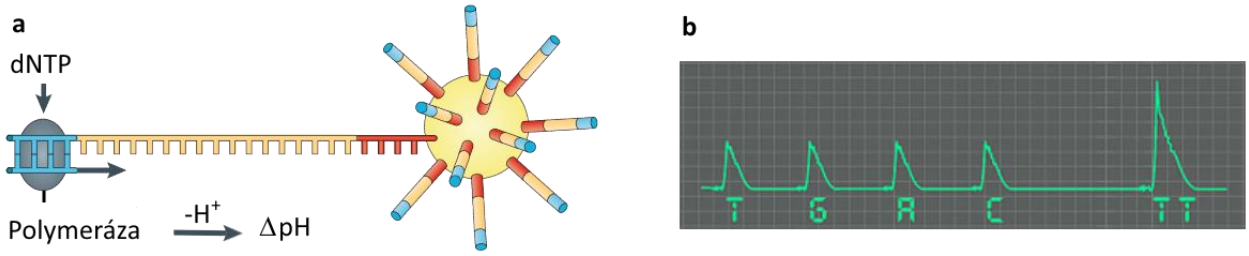
Obrázek 5: SOLiD sekvenování – a) ligace próby k primeru, excitace a detekce fluoroforu, štěpení za pátým nukleotidem a opakování ligačních cyklů. Následuje resetování primeru a další kolo sekvenování s primerem kratším o jednu bázi; b) schéma zarovnání sekvenovaných molekul a dekódování sekvence pomocí barevného schématu (převzato a upraveno z Metzker, 2010).

2.1.4 Ion Torrent (Life Technologies)

Ion Torrent, dnes patřící také pod firmu Life Technologies, používá podobnou sekvenační strategii jako 454 pyrosekvenování (Glenn, 2011), ale místo pyrofosfátu je zde detekován vodíkový ion, který je odštěpován jako druhý vedlejší produkt při syntéze DNA. Výhodou metody je to, že nejsou potřeba žádné kamery ani fluorofory – tato technologie přímo převádí chemicky kódovanou informaci (pořadí bází) do formy digitální informace (0, 1) na polovodičovém čipu (www.lifetechnologies.com[1]).

K amplifikaci sekvenované DNA se používá opět emulzní PCR (Dressman et al., 2003) fragmentů s naligovanými adaptory. Po navázání primerů a polymerázy na templát se vzorek

napipetuje na čip a centrifugací jsou mikro kuličky umístěny do jednotlivých jamek. I zde je velikost jamky zvolena tak, aby se do ní vešla právě jedna mikro kulička (Rothberg et al., 2011). Po začlenění nukleotidu do vznikajícího řetězce a odštěpení vodíkového iontu dojde ke změně pH roztoku, což je detekováno senzorem jako změna napětí. V případě začlenění více stejných nukleotidů za sebou, bude změna pH dosahovat příslušných násobků hodnoty pro jeden nukleotid (Obrázek 6; www.lifetechnologies.com[1]). Jamkami postupně proudí všechny nukleotidy, mezi kterými je zařazen promývací krok, aby se zabránilo ulpívání předešlých nukleotidů v jamkách. Přesnost Ion Torrent je vyšší, než u optických metod používajících modifikované nukleotidy, nejčastější chyby jsou pak spojeny s homopolymerními úseky (Rothberg et al., 2011).



Obrázek 6: Ion Torrent – a) při inkorporaci nukleotidu se jako vedlejší produkt odštěpuje ion vodíku, který sníží pH v sekvenační jamce; změna pH je detekována na polovodičovém čipu jako změna napětí. Při inkorporování dvou stejných nukleotidů po sobě je naměřené napětí dvojnásobné (převzato a upraveno z Metzker, 2010; www.lifetechnologies.com[1]).

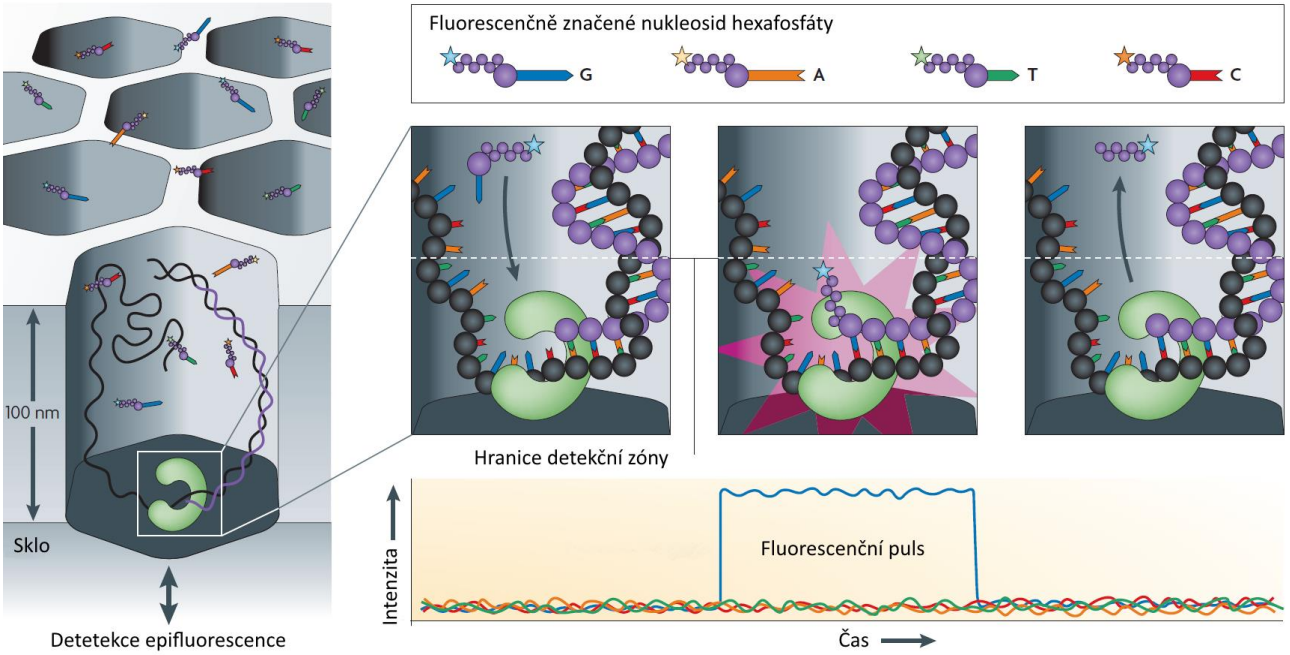
2.1.5 PacBio (Pacific Biosciences)

Metoda sekvenování firmy Pacific Biosciences nazvaná PacBio (www.pacific-biosciences.com) patří k metodám sekvenujícím jednotlivé molekuly (tzv. „single-molecul sequencing“) – bez nutnosti je předem amplifikovat, v reálném čase (tzv. „real-time“), bez přerušení syntézy (Obrázek 7; Eid et al., 2009).

Na dně každé sekvenační jamky se nachází imobilizovaná polymeráza, ke které difundují jednotlivé nukleotidy. Každý z nich je značen fluoroforem s jiným emisním spektrem. Excitační laserový paprsek směřuje do spodní části jamky (www.pacificbiosciences.com), kde detekuje nukleotid nacházející se v aktivním místě polymerázy. Fluorofor je vázán na terminálním fosfátu nukleosid hexafosfátu (Eid et al., 2009), díky čemuž nedochází ke sterickému bránění syntézy. Po štěpení nukleotidu je fluorofor uvolněn, a tím vzniká nemodifikovaný řetězec DNA (Kumar et al., 2005). Chybám této metody dominují delece, které mohou vznikat při příliš rychlém zabudování dvou nukleotidů po sobě – čas mezi dvěma fluorescenčními pulzy není dostatečně dlouhý pro spolehlivou detekci. Inzerce jsou pak způsobeny disociací nukleotidu z aktivního místa před tvorbou fosfodiesterové vazby, a tím zdvojením signálu (Eid et al., 2009).

Díky polymerázové kinetice citlivé k biologickým narušením mohou být pomocí platformy PacBio vyšetřovány např. i DNA vazebné proteiny, inhibitory DNA polymerázy nebo účinky

metylace. Vysoká délka readů (osekvenovaných fragmentů) je dosažena tím, že nedochází k zastavování syntézy DNA. K odstranění chyb spojených se sekvenováním jednotlivých molekul může být použit cirkulární templát, který je osekvenován několikrát za vzniku konsenzuální sekvence (Eid et al., 2009).



Obrázek 7: PacBio – na dně sekvenační jamky je imobilizovaná polymeráza, která syntetizuje komplementární řetězec k templátu. Fluorescenčně značené hexafosfátové nukleotidy jsou detekovány laserovým paprskem v aktivním místě polymerázy těsně před svým začleněním. Při inkorporaci je příslušný nukleotid štěpen a vzniká nemodifikované vlákno DNA (převzato a upraveno z Metzker, 2010).

2.2 Srovnání NGS platforem a jejich výhody a nevýhody oproti Sangerově metodě

Jednotlivé metody next-generation sekvenování se liší v mnoha ohledech. K nejdůležitějším z nich patří délka readů, počet readů za běh (viz Tabulka 1), cena jak za sekvenační běh, tak za megabázi dat a chybovost (shrnuto v Tabulce 2). Délka jednotlivých readů je důležitá, zvlášť pro tzv. „*de novo* assembly“, protože čím delší ready jsou, tím jednodušší je jejich spojení do delších sekvencí, kontigů. V délce readů vyniká PacBio s 20 000 bází (b), následuje Roche 454 s 1 000 b (van Dijk et al., 2014), klasická Sangerova metoda s 650 b a Illumina s 2× 300 bázemi sekvenovanými z obou konců. Naopak spojení readů SOLiD mající pouze 110 b je značně náročné oproti ostatním metodám (Glenn, 2014). Počet readů za sekvenační běh spolu s délkou readu ovlivňují cenu za gigabázi výsledných dat. Všechny metody next-generation sekvenování předčí Sangerovu metodu o několik řádů, nejvíce vyniká Illumina, SOLiD a Ion Torrent. Co se týká ceny sekvenování, je Sangerova metoda vhodná pouze pro velmi malé studie, protože gigabáze dat stojí mnohonásobně více, než u kterékoliv z NGS platforem. Nejnižší cenu za osekvenovanou gigabázi dat poskytuje Illumina, po ní následuje SOLiD, které má však vysokou cenu za sekvenační běh,

a Ion Torrent. PacBio disponuje nejnižší cenou za běh, cena gigabáze dat je však kvůli malému množství osekvenovaných readů v jednom běhu vyšší. Nejdražší metodou v přepočtu na gigabázi dat je Roche 454, které má i jednu z nejvyšších cen za jeden běh (Glenn, 2014).

Tabulka 1: Srovnání next-generation sekvenovacích metod a Sangerova sekvenování – druh amplifikace charakteristiky běhu (převzato a upraveno z Glenn, 2014 a van Dijk et al., 2014).

Název	*Amplifikace*	*Délka běhu*	*Délka readu*	*Počet readů/běh*	*Gb/běh*
Sangerovo sekv.	PCR	2 hod	650 b	96	$62{,}4 * 10^{-6}$
Roche 454	emulzní PCR	10 - 20 hod	400 - 1 000 b	$0{,}1 - 1 * 10^{6}$	0,05 - 0,7
Illumina	bridge PCR	5 h - 14 dní	36 - 600 b	$1 - 6\,000 * 10^{6}$	0,3 - 1 800
SOLiD	emulzní PCR	8 dní	110 b	$1410 * 10^{6}$	320
Ion Torrent	emulzní PCR	3 - 7 hod	175 - 400 b	$0{,}475 - 70 * 10^{6}$	0,095 - 87,5
PacBio	není	2 hod	20 000 b	$0{,}03 * 10^{6}$	0,5

V současnosti nejpoužívanější metodou NGS se stala Illumina, zejména díky kombinaci vysokého počtu readů za běh a nízké ceny. Druhou metodou je pak Roche 454, kterému v polovině roku 2016 skončí podpora, zejména kvůli finanční náročnosti analýz. PacBio trpí velkou chybovostí jednotlivých readů (Glenn, 2014), tudíž jako samostatný zdroj dat se příliš nehodí, nicméně lze ji uplatnit v kombinaci s jinou metodou s kratšími ready (Utturkar et al., 2014).

Tabulka 2: Srovnání ceny a chybovosti u next-generation sekvenovacích metod a Sangerova sekvenování. Chybovost u PacBio dosahuje této hodnoty však pouze v případě konsenzuální sekvence tří readů jednoho templátu (převzato a upraveno z Glenn, 2014).

Název	*Cena reagencií/běh*	*Cena reagencií / Gb*	*Nejčastější chyba*	*Chybovost*
Sangerovo sekv.	3 600 Kč	57 692 000 Kč	substituce	0.1-1 %
Roche 454	24 500 - 155 000 Kč	238 000 - 488 000 Kč	indel	1 %
Illumina	13 250 - 449 500 Kč	175 - 44 150 Kč	substituce	~ 0.1 %
SOLiD	262 500 Kč	1 700 Kč	A-T bias	≤ 0.1 %
Ion Torrent	8 750 - 25 000 Kč	2 040 - 91 850 Kč	indel	~ 1 %
PacBio	2 500 Kč	27 775 Kč	indel	≤ 1 %

Většina platforem disponuje nabídkou různých sekvenačních přístrojů, které se liší v jednotlivých parametrech. Kromě využití příslušné platformy je tedy možno/nutno vybírat i konkrétní přístroj podle potřeb dané biologické otázky. Např. Illumina HiSeq X, jehož běh je nejdražší ze všech sekvenovacích platforem, má také nejvyšší počet readů za běh a tím nejnižší cenu za vzorek, avšak střední délku readů a běhu. Oproti tomu Illumina MiSeq s nejdelšími ready má velmi nízkou kapacitu a relativně vysokou cenu za gigabázi dat (Glenn, 2014).

3 Molekulární metody pro řešení fylogenetických otázek

Do klasických molekulárních technik běžně využívaných pro řešení fylogenetických otázek na různých taxonomických úrovních řadíme především amplified fragment length polymorphism (AFLP; Vos et al., 1995), mikrosatelity, též krátké tandemové repetice (SSR; Jarne and Lagoda, 1996) a Sangerovo sekvenování (Sanger et al., 1977). AFLP je restrikční metoda, která umožňuje generovat velké množství nezávislých jaderných markerů. Výhodou je, že pro její využití není nutná znalost nukleotidové sekvence (Vos et al., 1995), nevýhodou pak, že se jedná o anonymní dominantní markery, u nichž může docházet ke komigraci nehomologních fragmentů (O'Hanlon and Peakall, 2000; Vekemans et al., 2002). Mikrosatelity jsou kodominantní markery – umožňují zjistit alelické složení jedince. Jejich využití bylo omezeno nutností znát alespoň část sekvence kvůli designu primerů (Jarne and Lagoda, 1996), avšak pomocí NGS je nyní možné rychle a efektivně identifikovat velké množství mikrosatelitů. Klasické Sangerovo sekvenování je využitelné u blízce i vzdáleně příbuzných jedinců, avšak pro použití je také nutná znalost vyšetřované sekvence kvůli designu primerů (Small et al., 2004) a pro fylogenetické analýzy, které zahrnují větší množství jedinců a lokusů, se jedná o přístup značně finančně náročný (Glenn, 2014).

I když se náklady na sekvenování celých genomů pomocí NGS rychle snižují – díky rostoucím délkám readů jednotlivých platforem a zvyšujícímu se počtu sekvencí získaných během jednoho běhu, celogenomové sekvenování zahrnující větší počet jedinců se zatím uplatňuje zejména u modelových druhů a druhů s malými genomy – z rostlin např. u huseníčku rolního (*Arabidopsis thaliana* L.; Schneeberger et al., 2011) nebo sóji luštinaté (*Glycine max* L.; Lam et al., 2010). Pro nemodelové druhy a studie zabývající se rostlinami s velkými genomy je tento přístup stále finančně velmi náročný, a proto bylo vyvinuto mnoho metod na přípravu DNA knihovny omezujících komplexitu genomu, např. amplikonové sekvenování (Meyer et al., 2008; Bybee et al., 2011), metody založené na restrikci DNA (Baird et al., 2008; Van Tassell et al., 2008), cílené obohacení (Gnirke et al., 2009; Okou et al., 2007), sekvenování transkriptomů (Morin et al., 2008; Marioni et al., 2008) nebo mělké sekvenování (Straub et al., 2012). Snížením komplexity genomu o jeden až dva řády se získají stovky až tisíce lokusů vhodných pro další analýzy (Van Tassell et al., 2008) s mnohem nižšími náklady za bázi, než u Sangerova sekvenování (Glenn, 2014).

3.1 Nejrozšířenější NGS metody a postupy pro přípravu DNA knihovny

3.1.1 Obecný postup přípravy DNA knihovny pro next-genetation sekvenování

Termínem DNA knihovna pro next-generation sekvenovací techniky se obecně označuje kolekce DNA fragmentů genomu určitého organismu nebo jejich skupiny, který je sekvenován

jednou z dostupných NGS platforem. Různé postupy přípravy DNA knihovny se od sebe značně liší, základní princip a některé kroky jsou však pro všechny metody společné.

Za prvé je nutné vybrat úseky, které budou osekvenovány. V některých případech je výběr fragmentů proveden náhodně (Straub et al., 2012), jindy je cíleno na přepisované oblasti (Morin et al., 2008; Bi et al., 2012) nebo organelární genomy (Parks et al., 2009; Stull et al., 2013). Za druhé se extrahovaná DNA musí nafragmentovat. To je možné provádět například restrikčním enzymem (Van Tassell et al., 2008; Elshire et al., 2011), sonikací (Straub et al., 2012; Morin et al., 2008), amplifikací pomocí PCR (Binladen et al., 2007; Meyer et al., 2008) nebo kombinací výše zmíněných postupů (Baird et al., 2008). Za třetí se na konce fragmentů přidají sekvenační adaptory s tagy (též často označované jako barcody) – krátkou specifickou nukleotidovou sekvencí, která umožňuje rozlišení jednotlivých vzorků při analýze dat po sekvenování knihoven několika až několika desítek spojených vzorků. Nejčastěji se používá ligace adaptorů (např. Baird et al., 2008), popř. mohou být přidány v průběhu PCR (Bybee et al., 2011). Za čtvrté následuje samotné sekvenování na vybrané NGS platformě. Pořadí některých kroků se může lišit v závislosti na konkrétní metodě přípravy knihovny nebo podle potřeb vybrané sekvenční platformy.

3.1.2 Amplikonové sekvenování a další metody založené na PCR

Amplikonové sekvenování, někdy zvané též „parallel tagged sequencing“, je obdobou klasického sekvenování. Stejně jako u Sangerovy metody dochází ke zmnožení cílových úseků DNA pomocí PCR, následuje však označení vzorků tagy a sekvenování mnoha spojených vzorků v jedné sekvenační reakci. Tag může být k DNA přidán při PCR amplifikaci pomocí primerů s tagem na 5´ konci, kde je potřeba před samotnou PCR nasyntetizovat primery s různými 5´ konci (Binladen et al., 2007), ligací specifických adaptorů obsahujících tag (Meyer et al., 2008) nebo v druhém kole PCR. Při této metodě jsou připraveny dva typy primerů, první pár je specifický pro amplifikovaný lokus, prodloužený o 5´ univerzální adaptor a je použit pro všechny vzorky. Druhý pár je vybrán ze sady primerů, které se skládají ze specifického tagu a adaptoru. Při první PCR je tedy příslušný lokus namnožen a v druhé PCR reakci označen (Bybee et al., 2011). Metodu založenou na ligaci je možno použít nejen na sekvenování celých PCR produktů. Po naamplifikování dlouhé sekvence je zařazen štěpící krok a až na vzniklé fragmenty se naligují specifické adaptory (Meyer et al., 2008), což snižuje počet nutných PCR reakcí, které jsou obecně zdrojem chyb kvůli chybovosti polymerázy a vzniku PCR chimér (Cronn et al., 2002). Protokol založený na dvou kolech PCR se naopak vyhýbá ligačním a purifikačním krokům, čímž se zkrátí čas na přípravu DNA knihovny a její cena (Bybee et al., 2011).

Výhodou amplikonového sekvenování jsou data s malým množstvím chybějících údajů (O'Neill et al., 2013; Griffin et al., 2011), nevýhodou pak nutnost provést PCR pro všechny lokusy u každého jedince zvlášť. Tento problém se snaží řešit multiplex PCR, metoda využívající více párů primerů pro amplifikaci různých lokusů v jediné reakci (Chamberlain et al., 1988), která má ovšem také značné množství nevýhod, jako je např. nerovnoměrné zastoupení produktů, amplifikace nechtěných oblastí nebo problémy s reprodukcí výsledků (Markoulatos et al., 2002). Tyto problémy se zvětšují s rostoucí fylogenetickou hloubkou např. kvůli rozdílné délce amplifikovaných lokusů a variabilitě v místech nasedání primerů (Lemmon and Lemmon, 2013).

Další možností, jak se vyhnout přípravě velkého množství amplifikačních reakcí, je použití mikrodroplet PCR. Tato metoda využívá pikolitrových kapek k amplifikaci mnoha cílových lokusů samostatně bez přítomnosti jiných primerů, ale současně v jedné reakci. Emulze vodných mikrokapek v oleji obsahující specifické dvojice primerů jsou smíchány s druhou sadou mikrokapek obsahující fragmenty genomové DNA a PCR reagencie. Po amplifikaci je olej odstraněn, vzorky jsou čištěny a sekvenovány. Mezi výhody patří vysoká reprodukovatelnost, vznik obdobného množství všech produktů a možnost amplifikace tisíců lokusů v jedné reakci (Tewhey et al., 2009a), což značně šetří čas a množství využitých reagencií, avšak jako nevýhoda se jeví vysoká cena za naamplifikovaný vzorek (Lemmon and Lemmon, 2013).

Amplikonové sekvenování je vhodnou metodou jak pro rekonstrukci hluboké fylogeneze (Bybee et al., 2011), tak pro fylogenezi nedávno oddělených druhů (O'Neill et al., 2013; Parks et al., 2009). Lze ji použít pro sekvenování celých organelárních genomů, jak chloroplastových (Parks et al., 2009; Njuguna et al., 2010), tak živočišných mitochondriálních (Chan et al., 2010). Mitochondriální rostlinné genomy metodou amplikonového sekvenování zkoumány dosud nebyly. Práce využívající tyto metody jsou založeny na menším množství lokusů a jedinců (Griffin et al., 2011; O'Neill et al., 2013), což je zřejmě způsobeno finanční a časovou náročností při použití na větší počet jedinců nebo lokusů. Samotná metoda amplifikace pomocí PCR skýtá také jisté nevýhody, jako jsou problémy s ekvimolárním sdružením jednotlivých vzorků (Binladen et al., 2007), PCR bias (přednostní amplifikace kratších alel nebo různé výtěžky cílů obsahující různé množství GC párů bází; Mutter and Boynton, 1995), možnost vnesení chyb polymerázou a vznik chimérických sekvencí (Cronn et al., 2002).

Griffin et al. (2011) se snaží metodu optimalizovat také pro polyploidní druhy rostlin. Místo jedné PCR pro vybrané jaderné lokusy lipnic (*Poa* L.) slučuje tři reakce, aby se zabránilo preferenční amplifikaci některé z alel. Na označení jednotlivých vzorků používá ligační metodu k navázání sekvenačních adaptorů s tagem na 3´ konci (Lennon et al., 2010). Jednotlivé genové kopie následně rozlišuje při analýze dat (Griffin et al., 2011). Amplikonové sekvenování bylo

využito také k rekonstrukci fylogeneze polyploidních jahodníků (*Fragaria* L.; Njuguna et al., 2010) a pelyňku (*Artemisia tridentata* Nutt.; Richardson et al., 2012), kde zřejmě došlo k selhání některých PCR reakcí nebo sekvenování, protože několik lokusů nebylo v datovém souboru zastoupeno.

3.1.3 Metody založené na restrikci DNA

K nejdůležitějším metodám redukce genomu založených na restrikci DNA patří restriction-site associated (RAD) sekvenování (Baird et al., 2008), reduced representation library (RRL; Altshuler et al., 2000), a genotyping by sequencing (GBS; Elshire et al., 2011), jejichž společným znakem je naštěpení genomové DNA na fragmenty pomocí restrikčního enzymu a jejich následné sekvenování jednou z platforem NGS.

V metodě RLL je mezi štěpení DNA a sekvenování vložen krok výběru velikosti vzniklých fragmentů. Ty se rozdělí na agarózovém gelu, následuje vyříznutí a izolace fragmentů o vhodné délce (Altshuler et al., 2000). Jako metodu pro přípravu NGS knihovny ji poprvé využil Van Tassel et al. (2008), čímž odpadá krok klonování fragmentů před sekvenováním.

Při RAD sekvenování je vzorek naštěpen, na oba konce je naligován adaptor P1, který se skládá z forward primeru, Illumina sekvenčního primeru a specifického tagu. Vzorky jsou spojeny, náhodně naštípány na kratší fragmenty a je vyizolována cílová velikostní frakce (Baird et al., 2008). Následuje ligace druhého adaptoru, který má tvar Y tvořený nepárujícími 5´ konci (Coyne et al., 2004) a sloužící jako reverse primer, díky kterému jsou při PCR preferenčně namnoženy fragmenty obsahující na jedné straně P1 a na druhé P2 adaptor. Oproti RLL jsou u RAD sekvenování získány sekvence z regionů obklopujících restrikční místa dané endonukleázy bez ohledu na délku fragmentu (Baird et al., 2008).

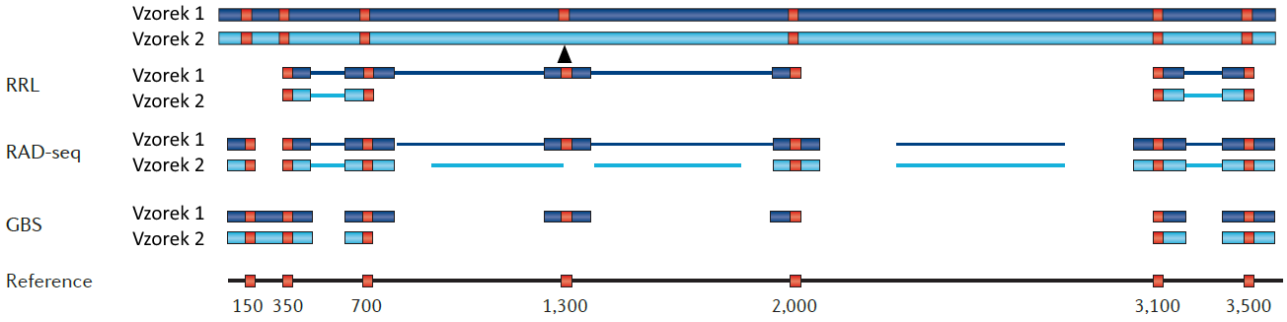
Metoda GBS byla vyvinuta pro druhy s velkým množstvím repetitivních sekvencí. Používá pro omezení genomové složitosti restrikční enzymy citlivé k metylacím a cílí tak štěpení, a tím i sekvenované oblasti, do míst s nízkým počtem kopií. Po ligaci adaptorů, které jsou různé pro oba konce templátové DNA, a spojení vzorků jsou fragmenty naamplifikovány a následuje samotná sekvenace. GBS je jednodušší oproti ostatním restrikčním metodám, nevyžaduje sonikaci ani výběr velikosti fragmentů. Nevýhodou naopak může být nerovnoměrné pokrytí u sekvencí s vyšším zastoupením menších fragmentů (Obrázek 8; Elshire et al., 2011).

RAD sekvenování bylo vytvořeno jako nástroj pro generování velkého množství jednonukleotidových polymorfismů (SNP) a polymorfismů v restrikčních místech (Baird et al., 2008). SNP, definované jako jednonukleotidové změny v DNA, jsou nejrozšířenějším typem sekvenční variability (Batley, 2003; Sachidanandam et al., 2001), čehož je využíváno

ve fylogenetických analýzách (např. Lu et al., 2013). Mutace v restrikčních místech naopak nejsou pro fylogenetiku vhodné, mohou vést k velkému počtu chybějících údajů (Lu et al., 2013; Rubin et al., 2012), a tím ovlivňovat i následné analýzy. Se vzrůstající fylogenetickou vzdáleností jedinců se značně snižuje počet vzniklých homologních fragmentů, a proto jsou metody založené na restrikci DNA vhodné pouze pro fylogenetické otázky na nižší úrovni – u blízce příbuzných jedinců a rychle se vyvíjejících druhů (Rubin et al., 2012).

Zvýšit počet lokusů zahrnutelných do fylogenetické analýzy, a tím zlepšit účinnost a využitelnost RAD sekvenování, se snaží double-digest RAD. Oproti původní metodě je krok náhodného štěpení fragmentů nahrazen restrikční reakcí obsahující dva enzymy – jeden s vzácnými štěpícími místy a jeden s běžně rozpoznávanou sekvencí v genomu. Použitím double-digest RAD sekvenování nedochází k náhodnému štěpení delších fragmentů, a tudíž náhodnému výběru následně sekvenovaných fragmentů, ale k současnému štěpení druhým enzymem, čímž je zajištěn výběr více korelujících fragmentů napříč jednotlivými vzorky (Peterson et al., 2012).

Výběr restrikčních enzymů značně ovlivňuje počet vzniklých fragmentů, např. u double-digest RAD sekvenování může být dosaženo generování od stovek do stovek tisíc lokusů (Peterson et al., 2012). Také je možné se vyhnout repetitivním oblastem využitím restrikčních enzymů citlivých na metylace (van Orsouw et al., 2007; Elshire et al., 2011), protože repetitivní sekvence jsou oproti genům často metylované (Rabinowicz et al., 1999). Mezi výhody metod založených na restrikci DNA patří časová nenáročnost přípravy DNA knihovny (Baird et al., 2008), vysoká účinnost a nízká cena (Van Tassell et al., 2008; Cronn et al., 2012), markery generované rovnoměrně z celého genomu, ne z jednoho konkrétního místa, a také to, že pro jejich využití nejsou potřeba žádné informace o sekvenci fragmentů, což je vhodné pro výzkum nemodelových druhů bez vlastních nebo blízkých referenčních sekvencí.



Obrázek 8: Srovnání osekvenovaných oblastí tří metod namapovaných na referenci – RLL, RAD a GBS. Nahoře jsou původní vzorky DNA, u vzorku 2 chybí restrikční místo v 1300 b (označeno šipkou); u RLL nejsou do sekvenování zahrnuty příliš velké ani malé fragmenty, stejně jako konce DNA bez restrikčního místa pro navázání adaptoru. RAD fragmenty jsou náhodně štěpené, proto je zahrnut fragment i mezi 700 a 2000 b u vz. 2. Tenké linky označují získané sekvence při sekvenování fragmentů i z konce neobsahující restrikční místo. GBS sekvenuje krátké fragmenty včetně koncových částí DNA díky dvěma různým adaptorům ligovaným na DNA (převzato a upraveno z Davey et al., 2011).

Metody založené na restrikci DNA byly u polyploidů využity při zjišťování evolučních vztahů několika desítek populací prosa (*Panicum virgatum* L.; Lu et al., 2013) a rekonstrukci

fylogeneze dvou druhů temperátních bambusů (Wang et al., 2013). Genom 840 jedinců prosa byl vyšetřován metodou GBS, která byla vybrána kvůli velkému genomu a chybějící referenci u tohoto druhu. Sekvenování bylo provedeno s nízkým pokrytím, což vedlo k velkému množství chybějících údajů, avšak bylo prokázáno, že v případě využití velkého množství SNP, neovlivní tato chybějící data fylogenetické analýzy (Lu et al., 2013).

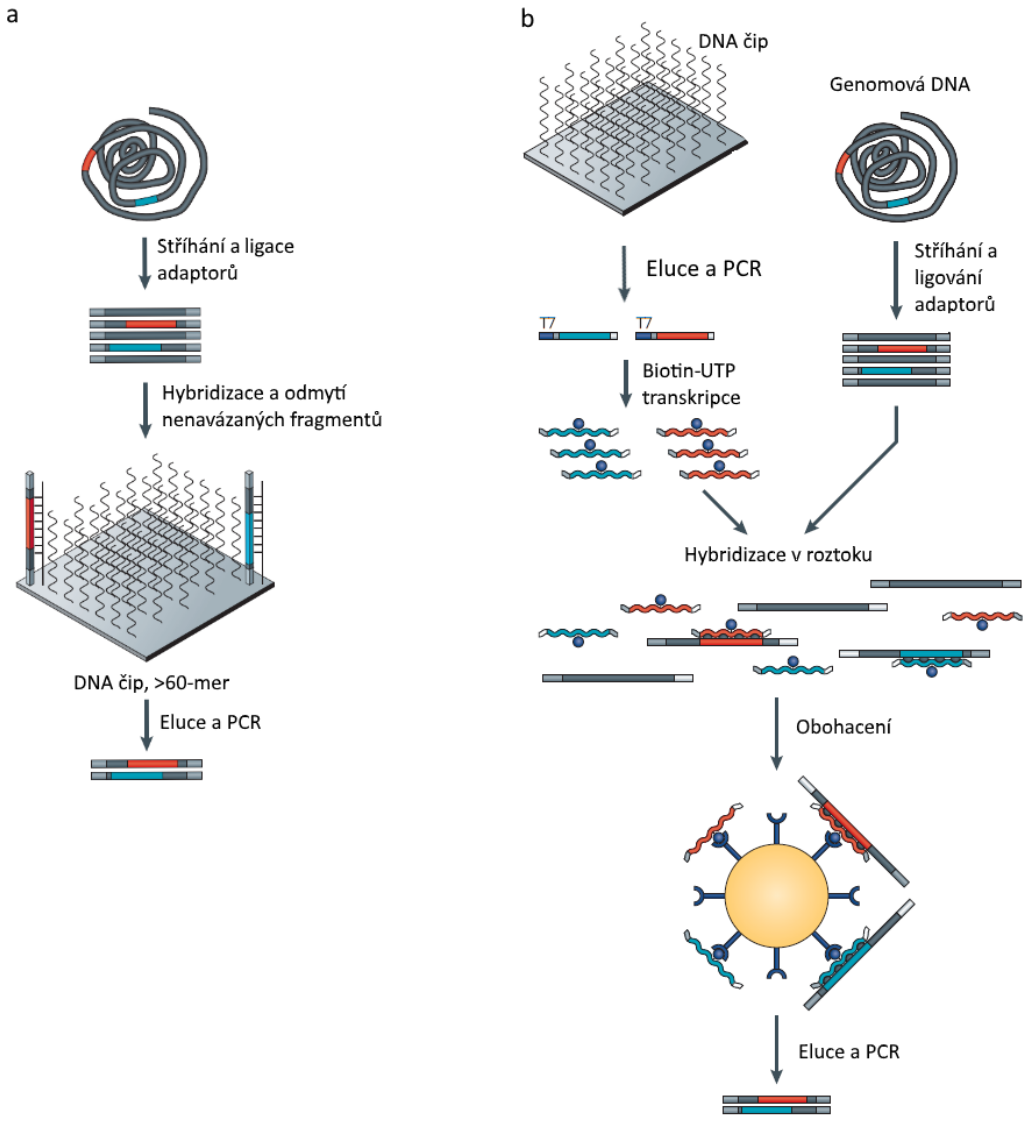
3.1.4 Cílené obohacení („target enrichment“)

Jak už název napovídá, metody cíleného obohacení (též někdy zvané jako „sequence capture“ nebo „targeted resequencing“) zachycují a obohacují cílové části genomu, a to buď hybridizací s DNA nebo RNA sondami – 60-120 bp (párů bází) dlouhými oligonukleotidy upevněnými na čipu („solid-phase capture“; Okou et al., 2007) nebo rozptýlených v roztoku („solution-phase capture“; Gnirke et al., 2009). K cílenému obohacení patří mnoho metod, např. PCR bait capture (Maricic et al., 2010), primer extension capture (PEC; Briggs et al., 2009), molecular inversion probe (MIP; Porreca et al., 2007; Turner et al., 2009), exonové obohacení (Bi et al., 2012; Weitemier et al., 2014), obohacení ultrakonzervovaných elementů (UCE; Faircloth et al., 2012), organemární obohacení mitochondriálního genomu (Briggs et al., 2009; Maricic et al., 2010) nebo plastidového genomu (Cronn et al., 2012; Stull et al., 2013).

Oproti obohacení na čipu, kde se hybridizují fragmenty DNA s DNA sondami připevněnými jedním koncem k podložce (Okou et al., 2007), se k obohacení v roztoku používají biotinylované RNA sondy. Ty jsou po hybridizaci s příslušným vzorkem DNA zachyceny pomocí kuliček obalených streptavidinem, které váží biotin z RNA sond (Gnirke et al., 2009). U obou metod následuje odmytí zbytkové DNA, eluce (oddělení DNA od sond), amplifikace pomocí PCR a sekvenování (Obrázek 9). Účinnost obohacení cílových sekvencí se může značně lišit. U rostlinných jaderných lokusů mohou cílené sekvence zaujímat asi od 20% (obohacení 1800x; Fu et al., 2010) až do 60% (2900x) všech dat. U polyploidních rostlin dochází také k obohacení všech kopií paralogů s vysokou efektivitou (Saintenac et al., 2011). Účinnost zachycení jednotlivých lokusů ovlivňuje i obsah GC párů bází v sondách. Nejlépe obohacují sondy s 45 % GC bází, u 23 a 66 % GC párů je účinnost zhruba poloviční. Vhodná koncentrace hybridizačních sond je dvojnásobná oproti koncentraci cílových sekvencí, použití většího nadbytku sond již dále účinnost cíleného obohacení nezvyšuje (Tewhey et al., 2009b).

Metoda cíleného obohacení je vhodná v případě známého referenčního genomu (některého z vyšetřovaných, popř. příbuzných druhů), z jehož sekvence je možné provést design hybridizačních sond. Pro nemodelové druhy lze použít tzv. „transcriptome based exon capture“ – ze sekvencí transkriptomu jednoho jedince navrhnout sondy i pro ostatní zkoumané druhy.

Nevýhodou tohoto přístupu může být vysoké procento nenamapovaných readů, pravděpodobně kvůli nekompletnímu referenčnímu genomu. Limitací je také nepřítomnost sond pokrývajících rozhraní exonů a intronů a nižší pokrytí na okrajích kódujících oblastí (Bi et al., 2012). Výhodou cíleného obohacení je jednoduché cílení na nerepetitivní informativní lokusy v genomu a možnost použití i na degradované vzorky nebo herbářové položky (Briggs et al., 2009; Sousa et al., 2014). V porovnání s ostatními se jedná o relativně drahou metodu využitelnou spíše u větších projektů, protože s cílením na delší oblasti v genomu klesá cena za zachycenou bázi (Cronn et al., 2012).



Obrázek 9: a) cílené obohacení na čipu – DNA se nafragmentuje a přiligují se adaptory, poté se hybridizuje se sondami na DNA čipu. Po odmytí zbytkové DNA se cílové sekvence eluují a amplifikují pomocí PCR; b) cílené obohacení v roztoku – DNA sondy nasyntetizované na čipu se odštěpí a namnoží pomocí PCR. In vitro transkripcí se sondy upevněné na čipu přepíší do RNA a začlení se biotin-UTP. Po hybridizaci jsou RNA sondy s cílovou DNA vychytány magnetickými kuličkami se streptavidimen vážící se na biotin. Následuje eluce DNA fragmentů a PCR amplifikace (převzato a upraveno z Metzker, 2010).

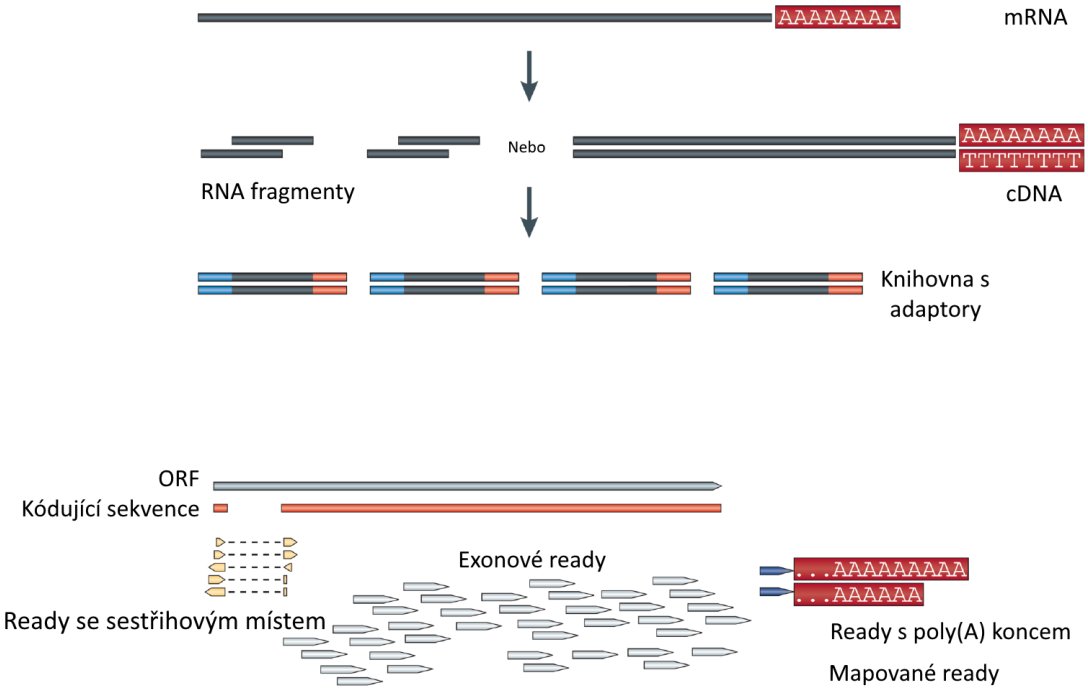
Cílené obohacení bylo použito u polyploidních druhů při sestavení fylogeneze jednotlivých chromozomů u oktapliodních jahod (*Fragaria* L.; Tennessen et al., 2014), chloroplastových genomů

několika desítek kultivarů brambor (*Solanum tuberosum* L.; Uitdewilligen et al., 2013) a posouzení vhodnosti znovuzavedení druhu *G. ekmanianum* Wittm. (Grover et al., 2015).

3.1.5 Sekvenování transkriptomu („RNA-seq“)

Sekvenování transkriptomu obecně sestává z několika kroků – extrakce ribonukleové kyseliny (RNA) z cílového organismu, reverzní transkripce do komplementární DNA (cDNA) a sekvenování cDNA knihoven (Obrázek 10; Morin et al., 2008; Marioni et al., 2008), čímž je docíleno osekvenování přepisované části genomu nazývané transkriptom, která tvoří od 1 % do 25 % celkového obsahu DNA (Cronn et al., 2012). Vzorek RNA může být osekvenován buď kompletní nebo jeho frakce, např. pouze polyadenylovaná mRNA izolovaná z kompletního vzorku oligo(dT) kuličkami (Marioni et al., 2008) nebo i nepolyadenylovaná frakce obohacená o neribozomální RNA (Morin et al., 2008). Pro fylogenetické studie je vhodné též zařadit krok normalizace DNA knihovny využívající duplex-specifickou nukleázu, jež sníží počet hojně zastoupených cDNA ve vzorku (Zhulidov et al., 2004).

Transkriptomové sekvenování bylo vybráno pro projekt The 1000 Plants (známý též jako oneKP nebo 1KP), jehož cílem je shromáždit více než 1000 rostlinných transkriptomů (www.onekp.com). Tato data mohou být využita nejen pro samotné fylogenetické analýzy, ale také při designu PCR primerů (Tonnabel et al., 2014) nebo hybridizačních sond pro metodu cíleného obohacení (Saintenac et al., 2011).



Obrázek 10: Transkriptomové sekvenování. Extrahovaná RNA je přepsána reverzní transkriptázou do cDNA, následuje fragmentace, ligování adaptorů a samotné sekvenování. Výsledkem jsou tři typy readů – s poly(A) koncem, exonové a obsahující sestřihové místo (převzato a upraveno z Wang et al., 2009).

Mezi výhody sekvenování transkriptomu se řadí jednoduchost cílení na přepisované a protein kódující oblasti, jejichž sekvenční informace nemusí být dopředu známa. Jde tudíž o přístup vhodný i pro druhy, kde dosud není k dispozici žádný, byť blízký referenční genom. Avšak nevýhod a problémů pojících se s touto metodou a jejím využití ve fylogenetice můžeme nalézt také několik. Jedním z nich je transkripce rozdílných genů v závislosti např. na typu orgánu či pletiva nebo vývojového stádia (Portnoy et al., 2011; Severin et al., 2010), což může být řešeno odebráním vzorků ze stejné tkáně ve stejné fázi životního cyklu (Bombarely et al., 2014) nebo spojením vzorků RNA z různých orgánů a/nebo vývojových stádií (Cavanagh et al., 2013). Další nevýhodou je potřeba čerstvého materiálu pro izolaci RNA a také samotná práce s RNA, která je náročnější – z důvodu vyšší náchylnosti k degradaci než DNA. Reverzní transkripcí do cDNA mohou být do sekvencí vneseny chyby, protože chybovost reverzní transkriptázy je vyšší, než je tomu u DNA polymerázy s proofriedingovou (opravnou) aktivitou (Roberts et al., 1989) a může také dojít ke vzniku chimérických sekvencí (Zeng and Wang, 2002). Značná variabilita v zastoupení jednotlivých RNA může vést ke ztrátě dat málo exprimovaných oblastí a polyploidních druhů může docházet k přednostní expresi homeologních genů jednoho rodičovského genomu (např. Bombarely et al., 2012; Hovav et al., 2008), a tím k nalezení menšího množství alel, než je přítomných v genomu (Bombarely et al., 2012). Vysoká cena za vzorek v porovnání s ostatními metodami pro přípravu DNA knihoven a náročnější assembly výsledných readů z důvodu alternativního sestřihu (Cronn et al., 2012) nepřispívá k využitelnosti metody ve fylogenetických rekonstrukcích polyploidních druhů.

Sekvenování transkriptomu bylo u polyploidních rostlin použito například u podčeledi agávovité (*Agavoideae*; McKain et al., 2012), při datování celogenomových duplikací v čeledi bobovité (*Fabaceae* Lindl.; Cannon et al., 2015), objevu množství SNP v genomech tolice vojtěšky (*Medicago sativa* L.; Li et al., 2012) nebo datování γ polyploidizační události (Jiao et al., 2012). Tato metoda nemusí být pro všechny fylogenetické otázky nejvhodnějším řešením, jak zjistili i ve studii zabývající se několika druhy tabáků (*Nicotiana* L.), kde okolo 70 % sestavených transkriptů postrádalo polymorfismy, což značně zredukovalo množství dat použitelných pro následné analýzy (Bombarely et al., 2012). Vzhledem ke značným nevýhodám a problémům transkriptomového sekvenování by bylo pro fylogenetické studie polyploidních rostlin zřejmě vhodnější využít genomová data namísto transktiptomových.

3.1.6 Mělké sekvenování („genome skimming“)

U mělkého sekvenování nedochází oproti základnímu konceptu přípravy DNA knihovny k žádným změnám – extrahovaná DNA je naštěpena a na oba konce fragmentů jsou naligovány

adaptory, avšak nedochází zde k žádné specifické selekci úseků, čímž se příprava knihovny výrazně usnadní. Následuje selekce velikosti, PCR obohacení, sloučení vzorků v ekvimolárním poměru a sekvenování. Pokrytí jednotlivých lokusů závisí na jejich relativním zastoupení v původním vzorku, to znamená, že dojde k hlubokému osekvenování DNA, která je přítomna v mnoha kopiích, jako je plastom, mitochondriální DNA a jaderná ribozomální DNA. Minimální hloubka pro kvalitní sestavení ribozomální DNA je 40x a chloroplastové DNA 30x (Straub et al., 2012). Část dat tvoří také sekvence z nízkokopiových jaderných lokusů, ve které je zastoupena velká část protein kódujícího genomu, čehož může být využito např. při vývoji PCR primerů (Straub et al., 2011) nebo hybridizačních sond (Cronn et al., 2012).

Nevýhodou této metody je komplikovaná optimalizace sekvenačního protokolu kvůli různému obsahu cílových sekvencí v jednotlivých vzorcích (Straub et al., 2012). Počet kopií ribozomální DNA na buňku se značně liší nejen mezi druhy, ale i na úrovni populací (Rogers and Bendich, 1987) a obsah chloroplastové DNA je různý mezi buňkami v jednotlivých orgánech a vývojových stádiích (Rauwolf et a l., 2010). Také může docházet ke značným nesrovnalostem mezi fylogenetickými stromy odvozenými od různých genomů – chloroplastového, mitochondriálního i jaderné ribozomální DNA (Straub et al., 2012; Bock et al., 2014). Naopak výhodou je jednoduchost přípravy a potřeba extrahovat pouze malé množství DNA, které může být získáno např. i z herbářových položek (Straub et al., 2012).

I když bylo mělké sekvenování použito ke zjištění původu hexaploidní slunečnice topinambur (*Helianthus tuberosus* L.; Bock et al., 2014) a fylogeneze jahodníků (*Fragaria* L.;Govindarajulu et al., 2015), jako vhodnější metody pro fylogenetické analýzy polyploidních druhů rostlin se jeví ty, jež využívají nízkokopiové jaderné lokusy, které nejsou na rozdíl od cytoplazmatických genomů děděny pouze uniparentálně (Small et al., 2004) a oproti ribozomální DNA netíhnou k homogenizaci homeologních lokusů (Wendel et al., 1995; Álvarez and Wendel, 2003). Ve studii Bock et al. (2014) bylo zvoleno mělké sekvenování z důvodu nehomogenizované ribozomální DNA a její vysoké fylogenetické informativnosti u zkoumaného rodu. Chloroplastová a mitochondriální DNA však byla značně variabilní, což je zřejmě způsobeno vícenásobným vznikem polyploidů a fylogenetické stromy vzniklé z těchto dat se značně lišily od stromu získaného z ribozomální DNA.

3.2 Srovnání NGS metod a jejich výhody a nevýhody oproti klasickým metodám

Různé metody přípravy DNA knihovny pro next-generation sekvenovací techniky se vzájemně liší v mnoha ohledech. Mezi nejdůležitější z nich patří cena a šíře aplikace, shrnuto v Tabulce 3. Nejlevnější je restrikční metoda GBS, nejdražší pak sekvenování transkriptomu a PCR

metody, jejichž cena se značně zvyšuje s rostoucí délkou cílové sekvence. Nejširší uplatnění má cílové obohacení a sekvenování transkriptomu, nejužší pak restrikční metody. Další z důležitých charakteristik je například vhodnost metody pro využití u nemodelových druhů, množství chybějících údajů, využitelnost u degradovaných vzorků nebo délka přípravy knihovny, shrnuto v Tabulce 4. Pro nemodelové druhy jsou vhodné restrikční metody, sekvenování transkriptomu a mělké sekvenování. Nejmenší množství chybějících údajů vzniká při cílovém obohacení, na degradované vzorky lze použít cílové obohacení a mělké sekvenování a krátkou dobou přípravy knihovny disponuje sekvenování transkriptomu a mělké sekvenování.

Tabulka 3: Srovnání ceny a šířky aplikací metod na přípravu DNA knihovny. Cena je počítána pro 96 vzorků a 50 nebo 500 kbp dlouhé cílové sekvence. „-" = metoda není vhodná pro danou aplikaci; „±" = metoda může být použita, ale existují i vhodnější metody; „+" = metoda je vhodná pro danou aplikaci (převzato a upraveno z (Cronn et al., 2012).

Metoda	*Zaměření experimentu*		*Přibližné náklady na vzorek (počet vykonaných reakcí)*	
	populace/ druhy	*druhy/ rody*	*50 kbp*	*500 kbp*
Krátká PCR (500 bp/amplikon)	±	+	3 000 Kč (9 600)	25 300 Kč (96 000)
Dlouhá PCR (5 000 bp/amplikon)	±	+	4 100 Kč (960)	6 600 Kč (9 600)
Restrikční metody - GBS	+	-	600 Kč (96)	1 000 Kč (96)
Restrikční metody - RAD	+	-	2 700 Kč (96)	3 100 Kč (96)
Cílené obohacení (synt. 2Mbp prób)	+	+	4 600 Kč (96)	4 600 Kč (96)
Sekvenování transkriptomu	+	+	8 300 Kč (96)	8 300 Kč (96)

Pro řešení otázek fylogeneze polyploidních druhů je vhodné využít metodu poskytující větší množství sekvencí za běh, např. Illumina, protože oproti diploidům je potřeba vyšší pokrytí k odhalení všech homeologních lokusů (Griffin et al., 2011). Také se zvyšující se délkou readu roste pravděpodobnost odlišení homeologních sekvencí (Clevenger et al., 2015). PacBio zatím vyřazuje vysoká chybovost, Roche 454 nebo Illumina se zdají být vhodné.

Metody přípravy knihoven pro sekvenování nové generace na jednu stranu vycházejí z klasických metod (a jsou některými svými charakteristikami velmi podobné), na druhou stranu využívají i zcela nové postupy, které ústí v kombinaci výhod i několika klasických metod v jedné NGS metodě. Např. metody založené na restrikci DNA – RADSeq (Baird et al., 2008), GBS (Elshire et al., 2011), RLL (Van Tassell et al., 2008) atd., jsou částečně podobné některým klasickým molekulárním technikám, jako je AFLP. Stejně jako AFLP nevyžadují NGS restrikční metody znalost sekvence zkoumaných druhů, avšak nastříhané fragmenty se neseparují elektroforézou podle velikosti, ale sekvenují a vyhledávají se polymorfismy v získaných

sekvencích, což je značná výhoda oproti původním metodám, kde nebylo jisté, zda stejně dlouhé proužky na skórovaném gelu jsou si homologní nebo pochází z odlišné části genomu. Markery již nejsou anonymní, čímž se lépe odlišují i homologní a homeologní lokusy.

Jiným příkladem je amplikonové sekvenování (Meyer et al., 2008; Bybee et al., 2011), které je obdobou klasické Sangerovy metody, nebo cílené obohacení (tzv. „target enrichment"; Gnirke et al., 2009; Okou et al., 2007), kde je možné v rámci jednoho sekvenačního běhu osekvenovat desítky, stovky i tisíce předem cíleně vybraných fragmentů u několika desítek i stovek jedinců, což značně kontrastuje s klasickým sekvenováním, kde v rámci jednoho sekvenačního běhu osekvenujeme maximálně 96 různých fragmentů.

Tabulka 4: Srovnání vybraných charakteristik metod na přípravu DNA knihovny. „-" = nesouhlasí s danou charakteristikou; „±" = částečně souhlasí; „+" = souhlasí s danou charakteristikou (převzato a upraveno z Lemmon and Lemmon, 2013 a Straub et al., 2012).

Metoda	*Vhodné pro nemodelové druhy*	*Minimum chybějících údajů*	*Lze použít na degradované vzorky*	*Krátká příprava DNA knihovny*
PCR metody	±	±	±	-
Restrikční metody	+	-	±	±
Cílené obohacení	±	+	+	-
Sekvenování transkriptomu	+	±	-	+
Mělké sekvenování	+	±	+	+

Hlavními výhodami next-generation sekvenování oproti klasické Sangerově metodě je mnohonásobně nižší cena za osekvenovanou bázi a generování obrovského množství dat v jediném běhu (Glenn, 2014). Také již není potřeba elektroforetických gelů a odpadá klonování fragmentů před samotným sekvenováním (např. Fortune et al., 2008; Rousseau-Gueutin et al., 2009), které je časově a finančně náročné. Klonování jednotlivých homo/homeologních alel je nahrazeno sekvenováním směsných vzorků a k jejich oddělení dochází až při analýze dat (Griffin et al., 2011).

Díky možnostem NGS sekvenování lze dnes do fylogenetických studií zahrnout stovky až tisíce jednotlivých lokusů (např. Cannon et al., 2015; Tennessen et al., 2014) oproti jednotkám až desítkám využívaným při Sangerově sekvenování (např. Fortune et al., 2008; Rousseau-Gueutin et al., 2009). Vysoké množtví získaných dat a informací z různých částí genomu může pomoci při řešení složitých příbuzenských vztahů, avšak jiným úskalím může být analýza tak velkého množství dat, která se jeví stále problematickou. Analýza NGS dat totiž není jen velmi náročná na hardware, software a čas strávený analýzou, ale i na absenci teoretických modelů a hypotéz, které slouží jako základní podklady pro jakoukoli analýzu dat (např. Pickrell and Pritchard, 2012).

Také se zjednodušilo sekvenování nemodelových druhů, tedy těch, které neměly žádnou ani blízce příbuznou referenční sekvenci. Oproti Sangerově sekvenování totiž NGS nevyužívá sekvenčně specifické primery, ale jednotné sekvenační primery, které se při přípravě DNA knihovny připojí k templátové DNA. Není proto nutná znalost jakékoliv, byť krátké genomové sekvence. Limitací Sangerovy metody byl náročný vývoj nových specifických primerů, proto se využívalo široce konzervativních sekvencí především z chloroplastového genomu nebo jaderné ribozomální DNA (rDNA; Sang, 2002; Small et al., 2004). Next-generation sekvenování, které nevyužívá specifických sekvenačních primerů, může jednoduše získávat sekvence z okolí restrikčních míst (Baird et al., 2008; Van Tassell et al., 2008), celého transkriptomu (Marioni et al., 2008; Morin et al., 2008), multikopiových oblastí (Straub et al., 2012) nebo vybraných cílových lokusů pro danou skupinu (Gnirke et al., 2009; Okou et al., 2007).

Pro fylogenetické analýzy polyploidních rostlin jsou využívány různé metody přípravy DNA knihoven, z nichž nejčastější je transkriptomové sekvenování. Ze sekvenovacích platforem bylo využito pouze 454 pyrosekvenování a Illumina, zřejmě z důvodu optimalizovaných protokolů a nízké ceny sekvenování metodou Illumina. 454 pyrosekvenování by však mohlo být nahrazeno i jinými metodami, např. Ion Torrent, které je oproti Roche 454 levnější (Glenn, 2014). Next-generation sekvenovací techniky byly dosud využity spíše při řešení fylogeneze na nižších úrovních, nejčastěji v rámci druhů jednoho rodu. Několik prací se zabývalo i hlubší fylogenezí, např. u *Agavoidae* nebo *Magnoliophyt*, jejichž data byla získána pomocí sekvenování transkriptomu (Tabulka 5).

Tabulka 5: Studie využívající NGS přístupy při rekonstrukci fylogeneze polyploidních rostlin

Studie	*Zkoumaný druh*	*Sekvenační platforma*	*Metoda přípravy DNA knihovny*
Griffin et al., 2011	*Poa* L.	Roche 454	amplikonové sekvenování
Cíl studie: zhodnocení využitelnosti NGS ve fylogenetice polyploidů s využitím 7 genů u 11 druhů *Poa* L.			
Hand et al., 2012	*Festuca arundinacea* Schreb.	Roche 454	amplikonové sekvenování
Cíl studie: sběr celogenomových SNP s využitím 414 lokusů u 12 morfotypů *F. arundinacea* Schreb.			
Njuguna et al., 2010	*Fragaria* L.	Illumina	amplikonové sekvenování
Cíl studie: odhalení sekvenčních rozdílů ve 22 druzích *Fragaria* L. s využitím 63 chloroplastových lokusů			
Richardson et al., 2012	*Artemisia tridentata* Nutt.	Roche 454	amplikonové sekvenování
Cíl studie: objasnění fylogenetických vztahů 3 podrodů *A. tridentata* Nutt. a zhodnocení sekvenční variability mezi podrody a ploidními úrovněmi pomocí 48 lokusů u 329 jedinců			
Arnold et al., 2015	*Arabidopsis arenosa* (L.) Lawalrée	Illumina	ddRAD (restrikční metody) a celogenomové sekven.
Cíl studie: rekonstrukce evoluční historie 20 populací *A. arenosa* (L.) Law.			

Tabulka 5: pokračování

Lu et al., 2013	*Panicum virgatum* L.	Illumina	GBS (restrikční metody)
Cíl studie: vývoj a analýza SNP pro vhled do evoluční dynamiky, fylogeneze, fylogeografie a populační struktury 6 ekotypů *P. virgatum* L.			
Wang et al., 2013	*Arundinaria* Michaux a *Yushania* P. C. Keng	Illumina	RAD (restrikční metody)
Cíl studie: identifikace SNP u 2 populací 2 druhů bambusů a hodnocení NGS pro fylogenezi bambusů			
Grover et al., 2015	*Gossypium* L.	Roche 454	cílené obohacení
Cíl studie: posouzení vhodnosti znovuzavedení druhu *G. ekmanianum* Wittm. s využitím 8 druhů *Gossypium* L.			
Tennessen et al., 2014	*Fragaria* L.	Illumina	cílené obohacení
Cíl studie: výzkum evolučních vztahů mezi subgenomy 2 oktaploidních a 5 diploidních druhů *Fragaria* L.			
Uitdewilligen et al., 2013	*Solanum tuberosum* L.	Illumina	cílené obohacení
Cíl studie: identifikace sekvenční variability 83 tetraploidních odrůd *Solanum tuberosum* L.			
Bombarely et al., 2012	*Nicotiana* L.	Roche 454	sekvenování transkriptomu
Cíl studie: objasnění evoluční historie 3 druhů rodu *Nicotiana* L. s využitím celogenomových markerů			
Bombarely et al., 2014	*Glycine* L.	Illumina	sekvenování transkriptomu
Cíl studie: pochopení původu a vývoje 3 polyploidních druhů *Glycine* L.			
Cannon et al., 2015	*Fabaceae* Lindl.	Illumina	sekvenování transkriptomu
Cíl studie: objasnění evoluční historie 37 druhů čeledi *Fabaceae* Lindl., datování polyploidizačních událostí			
Cavanagh et al., 2013	*Triticum aestivum* L.	Roche 454 a Illumina	sekvenování transkriptomu
Cíl studie: vývoj SNP a vytvoření SNP mapy, posouzení genetické variability 28 kultivarů *T. aestivum* L.			
Jiao et al., 2012	*Magnoliophyta*	Illumina	sekvenování transkriptomu
Cíl studie: datování γ polyploidizační události s využitím 35 druhů *Magnoliophyt*			
Li et al., 2012	*Medicago sativa* L.	Illumina	sekvenování transkriptomu
Cíl studie: identifikace SNP u 27 genotypů *M. sativa* L. a posouzení sekvenční diverzity mezi genotypy			
McKain et al., 2012	*Agavoideae*	Illumina	sekvenování transkriptomu
Cíl studie: testování hypotézy původu chromozomální bimodality v paleopolyploidizační události			
Bock et al., 2014	*Helianthus* L.	Illumina	mělké sekvenování
Cíl studie: zjistit rodičovské druhy *H. tuberosus* L. a odhalit vztahy mezi 8 druhy *Helianthus* L.			
Govindarajulu et al., 2015	*Fragaria* L.	Illumina	mělké sekvenování
Cíl studie: zjistit souvislost mezi rozporuplnými cytoplazm. genomy a evoluční historií 6 druhů *Fragaria* L.			

4 Polyploidní speciace a rekonstrukce fylogeneze u polyploidních druhů rostlin

4.1 Polyploidní speciace

Vznik polyploidních druhů patří k důležitým evolučním procesům. Alespoň jednou celogenomovou duplikací prošla celá větev semenných rostlin, krytosemenné rostliny zmnožily svůj genom dokonce vícekrát (Jiao et al., 2011). Polyploid, jedinec s více než dvěma sadami chromozomů, může vzniknout různými způsoby – splynutím neredukovaných gamet v rámci jednoho druhu, takový jedinec je označován jako autopolyploid, nebo zdvojením chromozomů u hybridního jedince za vzniku allopolyploida. Podle doby vzniku můžeme polyploidy dělit také na paleopolyploidy, mezopolyploidy a neopolyploidy. Tyto ternímy byly v minulosti vnímány a vykládány různě. V současné době se jako paleopolyploidní označuje druh geneticky diploidní, který prošel jednou nebo více dávnými polyploidizačními událostmi a tato minulost je identifikována až na základě analýzy ortologních sekvencí. Mezopolyploid je definován jako druh s diploidním genomem a nízkým počtem chromozomů, ve kterém jsou stále rozpoznatelné rodičovské subgenomy a neopolyploid vznikl nedávnou polyploidizační událostí a má zvýšenou genomovou velikost, počet chromozomů a genových kopií (Mandáková et al., 2010).

Duplikované geny odvozené od různých rodičovských subgenomů (mající původ v hybridizační nebo polyploidizační události) jsou označovány jako homeology. Oproti tomu homologními jsou nazývány geny pocházející ze stejného subgenomu (Dufresne et al., 2014). Osud homeologních genů může být různý. Dělíme jej na čtyři hlavní kategorie – redundanci (přítomnost vzájemně zastupitelných homeologů), neofunkcionalizaci (vznik nové nebo modifikované funkce u jedné homeologní kopie), subfunkcionalizaci (rozdělení původní funkce mezi homeology) a nonfunkcionalizaci (ztráta jedné genové kopie), která je nejčastější (Rastogi and Liberles, 2005). Evoluce homeologů je u polyploidních druhů velmi dynamická a liší se jak mezi různými taxony, tak mezi různými druhy genů (Adams and Wendel, 2005). Tato povaha polyploidních genomů umožňuje vznik evolučních novinek a je důležitá pro úspěch polyploidních rostlin, ale také vytváří problémy při rekonstrukci evolučních historií těchto druhů (Brysting et al., 2011).

4.2 Rekonstrukce fylogeneze polyploidních druhů rostlin a její úskalí

Mnoho fylogenetických studií i na nízkých úrovních, kde se mohou objevit síťovité vztahy díky alopolyploidii, byla donedávna založena na vyšetřování chloroplastových genomů (cpDNA) a jaderné ribozomální DNA (Lott et al., 2009). Tyto markery jsou však nevhodné pro řešení síťovitých vztahů z důvodu uniparentální dědičnosti cytoplazmatických genomů (Small et al., 2004)

a tíhnutí biparentálně děděné ribozomální DNA k homogenizaci homeologních lokusů (Wendel et al., 1995; Álvarez and Wendel, 2003) a rekombinaci mezi rodičovskými lokusy, čímž vznikají chimerické sekvence (Álvarez and Wendel, 2003). Ani kombinace rDNA a cpDNA markerů nemusí být dostatečná pro odhalení fylogeneze polyploidních druhů v případě, kdy je rDNA homogenizovaná k mateřskému genomu a také u vyšších polyploidů než tetraploidů, u kterých jsou přítomny více než dva rodičovské subgenomy (Brysting et al., 2011). Oproti tomu biparentálně děděné jaderné geny jsou méně náchylné ke vzájemné evoluci (tzv. „concerted evolution"; Cronn et al., 1999) a jeví se jako vhodnější zdroj informací pro fylogenetiku polyploidních druhů, odvození jejich síťovitých vztahů a rodičovských linií (Álvarez and Wendel, 2003; Small et al., 2004). Nevýhody jako problematický vývoj nových primerů nebo klonování pro odlišení jednotlivých homeologních lokusů (Sang, 2002; Small et al., 2004) jsou u next-generation sekvenování překonány – vývoj specifických primerů ani klonování již není nutné.

Také počet lokusů využívaný ve fylogenetických analýzách je často nízký. Některé studie použily pouze jediný nukleární gen pro odvození historie polyploidních skupin (např. Smedmark et al., 2003; Pfeil et al., 2004). To však není dostatečné, protože jeden gen odráží vývoj pouze malé části genomu, která nutně nemusí kopírovat vývoj jiných částí genomu (Záveská, 2014). Aby mohly být odlišeny polyploidizační události od ostatních biologických procesů, jež mohou být zodpovědné za nejednotnou topologii různých genových stromů, je nutné použít pro fylogenetické studie větší množství lokusů (Linder and Rieseberg, 2004). Studie využívající klasické Sangerovo sekvenování odvozují fylogenetické vztahy často pouze z několika lokusů, čímž vznikají stromy s některými špatně vyřešenými nebo málo podpořenými větvemi (např. Syring et al., 2007). Oproti tomu díky vysokému výkonu metod next-generation sekvenování je nyní možno osekvenovat stovky až tisíce lokusů několika desítek jedinců v jediném sekvenačním běhu, čehož se začalo využívat i ve fylogenetice při vytváření vysoce podpořených stromů (např. Cannon et al., 2015; Bombarely et al., 2014).

Studium polyploidních druhů pomocí next-generation sekvenování má také několik specifik oproti diploidním druhům. Prvním z nich je nutnost sekvenovat vzorky s vyšším pokrytím než u diploidních jedinců, protože pro 95% šanci zachycení obou alel u diploidního druhu stačí pokrytí 6×, pro 95% pravděpodobnost zachycení všech čtyř alel u tetraploidního jedince je potřeba pokrytí 15× (Griffin et al., 2011) a až 48× pro správné genotypování autotetraploidů kvůli odlišení digenického simplexu (*aaab*) od digenického duplexu (*aabb*; Uitdewilligen et al., 2013). Také je vhodné využít alespoň 150 bp dlouhé ready, protože s rostoucí délkou readu roste i pravděpodobnost odlišení homeologních sekvencí od sebe (Clevenger et al., 2015).

Co se týká protoloku na přípravu knihovny pro next-generation sekvenování, bylo doporučeno využití protokolů s minimálním množstvím kroků amplifikace pomocí PCR. Tato metoda může do sekvenovaných fragmentů zanést chyby, které jsou posléze mylně identifikovány jako polymorfismy, což může působit problémy u druhů, které mají malou variabilitu uvnitř nebo mezi subgenomy (Clevenger et al., 2015).

Fylogenetika využívající data next-generation sekvenování je založená buď na identifikaci jednonukleotidových polymorfismů nebo vytváří fylogenetické stromy z celých sekvencí jednotlivých lokusů. Metoda využívající celé sekvence se hodí spíše pro otázky na vyšších fylogenetických úrovních (např. Griffin et al., 2011; Weitemier et al., 2014), oproti tomu SNP jsou využívány spíše na nižších úrovních, především na vnitrodruhové úrovni (např. Li et al., 2012; Hand et al., 2012). Identifikovat však SNP u polyploidů je oproti diploidním druhům náročnější kvůli nutnosti rozlišovat mezi homeologními (polymorfní pozice z jiných subgenomů) a alelickými SNP (polymorfní pozice v rámci subgenomu). Čím bližší si subgenomy jsou, tím těžší je tyto dva druhy SNP odlišit. Allopolyploidní druhy se často chovají jako diploidní s disomickým děděním, naopak autopolyploidní druhy mají subgenomy téměř identické, což značně znesnadňuje analýzu (Clevenger et al., 2015).

4.3 Možnosti analýzy NGS dat

Typické úpravy surových dat z next-generation sekvenování před samotnou rekonstrukcí fylogeneze zahrnují ořezání adaptorů a nekvalitních částí readů. I když může být tímto filtrováním vyřazena velká část dat, použití přísnějších kritérií pro filtrování často vede k lepší assembly (sestavení krátkých readů do delších sekvencí, kontigů). Následuje buď tzv. „reference-guided assembly“ (sestavení readů do kontigů mapováním readů na již existující referenční sekvenci, která může být shodná se sestavovanou sekvencí nebo jí pouze podobná) nebo „*de novo* assembly“ (metoda sestavení readů do kontigů bez použití referenční sekvence). „*De novo“* přístup je však často výpočetně náročnější (Weitemier et al., 2014).

V případě metody tvorby fylogenetických stromů na základě SNP následuje jejich vyhledávání a identifikace. V dnešní době existuje celá řada programů, které k tomu lze využít (viz např. McCormack et al., 2013). Značně se však od sebe liší svými algoritmy, rychlostí i výstupy, z čehož vyplývá, že použitý program by měl být pečlivě vybírán. Vhodné je provést určení polymorfních pozic v několika různých programech a z nich vybrat ty, jejichž výsledky se nejvíce shodují. K následným fylogenetickým analýzám se použije shodující se výstup z obou programů, protože výstupy z jednotlivých programů se mohou významně lišit jak v počtu nalezených SNP, tak v počtu stejných polymorfních pozic nalezených mezi jednotlivými programy. Protože různé

programy využívají různé algoritmy, je logické, že pro skupiny rostlin s různou historií vzniku budou vhodné jiné programy. Některé skupiny rostlin disponují zvýšenou variabilitou z důvodu křížení vzdálených jedinců, domestikace nebo vícečetného vzniku polyploidních druhů, což vede k větší vzdálenosti mezi subgenomy a jednoduššímu odlišování homeologní SNP od alelických. Naopak zvýšené množství repetitivních sekvencí může znesnadnit namapování readů k jednotlivým subgenomům (Clevenger et al., 2015).

Pro znázornění evolučních vztahů mezi různými taxonomickými skupinami jsou nejčastěji využívány fylogenetické stromy (Linder and Rieseberg, 2004). S rozmachem využívání multilokusových genomických dat k odvozování fylogenetických stromů se hlavní výzvou staly konfliktní topologie různých genových stromů. Ty mohou být způsobeny několika různými procesy, shrnuto v Maddison (1997) nebo Degnan and Rosenberg (2009), např. genové duplikace a ztráty (zmnožení jednotlivého genu a následný samostatný vývoj dvou kopií), incomplete lineage sorting (proces, při kterém některá z linií splývá dříve se vzdálenější linií, než s linií bližší, tj. sdílí stejnou mateřskou alelu se vzdálenější linií, místo s linií bližší), hybridizace (křížení jedinců různých druhů), rekombinace (výměna části DNA mezi lokusy) nebo horizontální přenos (přijetí DNA od nerodičovského jedince). Odlišit hybridizaci od ostatních procesů způsobujících nejednotnost mezi genovými stromy by mělo být možné na základě nalezení sady konfliktních genových stromů, které se vyskytují častěji, než by odpovídalo v případě jejich náhodného vzniku (Linder and Rieseberg, 2004).

Pro tvorbu stromů z next-generation dat stejně jako u Sangerova sekvenování mohou být využity metody jako maximum parsimony (Camin and Sokal, 1965), maximum likelihood (Edwards and Cavalli-Sforza, 1964 podle Edwards, 2009) a Bayesovská analýza (Rannala and Yang, 1996), pro vizualizaci retikulárních vztahů pak neighbor net algoritmus (Bryant and Moulton, 2004). Maximum parsimony analýza může být prováděna např. v programu PAUP (Swofford 2002), maximum likelihood v programu RAxML 8 (Stamatakis, 2014), Bayesovská analýza v MrBayes 3 (Ronquist and Huelsenbeck, 2003) nebo BEAST 1.7 (Drummond et al., 2012) a neighbour network v programu SplitsTree (Huson and Bryant, 2006). Limitující pro tyto metody je ale množství dat, které jimi budeme zpracovávat. Pro výrazně větší NGS datasety nicméně bylo vyvinuto několik alternativ těchto klasických postupů, např. program NexABP využívající tzv. „anchor based“ metodu (Roychowdhury et al., 2013) nebo program Co-phylog patřící k tzv. „word frequencies based“ metodám. Jedná se o „assembly free“ přístup, který pracuje s jednotlivými krátkými ready, aniž by bylo nutné je sestavovat do kontigů (Yi and Jin, 2013).

5 Rešerše o příkladové skupině polyploidních rostlin – rodu *Curcuma* L.

Rod *Curcuma* L. patřící do čeledi *Zingiberaceae* v dnešní době čítá okolo 120 druhů. Centrum rozšíření se nachází v jižní a jihovýchodní Asii, největší diverzita druhů je v Thajsku a Indii. Kurkumy jsou důležité i z hospodářského hlediska, některé druhy jsou využívány jako barvivo, koření, okrasné rostliny nebo pro medicínské účely (Škorničková, 2007). Rod se dělí na tři podrody – *C.* subg. *Curcuma* K. Schum., *C.* subg. *Hitcheniopsis* (Baker) K. Schum. (Schumann, 1904 podle Záveská et al., 2012) a *C.* subg. *Ecomata* Škorničk. & Šída f. (Záveská et al., 2012) a vyznačuje se velkou genetickou rozmanitostí způsobenou hybridizací a polyploidizací se síťovitými vztahy mezi jednotlivými druhy (Záveská et al., 2011; Záveská et al., 2012).

V minulých letech výzkum tohoto rodu zahrnoval práci snažící se o odhalení vztahů v rámci podrodu *Curcuma*, kde byly nalezeny dvě vývojové větve. Jedna linie zahrnuje vyšší polyploidy a některé hexaploidní druhy vykazující síťovité vztahy, v druhé linii se pak nachází ostatní hexaploidní druhy (Záveská et al., 2011). Další práce publikovaná na toto téma se zabývala stanovením hranice rodu *Curcuma* L., ke kterému byly přiřazeny i druhy kurkumám podobné, avšak dříve spadající do jiných rodů. Také byl popsán nový podrod *Ecomata*, do něhož byly zařazeny dvě molekulárně, avšak ne morfologicky rozlišitelné linie *Pierreana* a *Ecomata*. (Záveská et al., 2012). V práci Záveská et al. (manuscript) byly pomocí čtyř jaderných genů odvozeny vztahy mezi čtyřmi hlavními molekulárními liniemi, které zůstaly dříve nevyřešeny (Záveská et al., 2012). Jako sesterské skupiny jsou považovány *Curcuma* s *Hitcheniopsis* a pak *Ecomata* s *Pierreana.* Na základě různé pozice v různých stromech (cpDNA, ITS, jaderné lokusy) byla navržena a testována hypotéza, že *C. vamana, C. candida, C. roscoeana* a *C. myanmarensis* jsou hybridního původu s rodiči pocházejícími z různých podrodů, což bylo u *C. vamana, C. roscoeana* a *C. myanmarensis* potvrzeno (Záveská et al., manuscript).

Některé otázky fylogeneze rodu *Curcuma* zůstávají stále nevyřešeny, například ačkoliv byly odhaleny síťovité struktury mezi vyššími polyploidy, ke zjištění jejich rodičovských druhů by mělo být využito větší množství (alespoň desítky) nezávislých genů (Záveská, 2014). I když hlavní vývojové skupiny jsou dobře podpořeny, vztahy mezi druhy uvnitř skupin jsou považovány za nevyřešené (Záveská et al., 2012) a i ke stanovení původu druhů *C. vamana, C. candida, C. roscoeana* a *C. myanmarensis* s vyšší pravděpodobností by mělo být využito větší množství markerů (Záveská, 2014).

5.1 Návaznost k diplomové práci

Ve své diplomové práci bych se chtěla zabývat fylogenezí polyploidního rodu *Curcuma* L. s využitím next-generation sekvenování. Chtěla bych ověřit vztahy mezi třemi podrody, původ meziskupinových hybridů navržených v práci Záveská et al. (manuscript), jakožto i dalších polyploidních druhů tohoto komplexu a také z dat získaných pomocí NGS pátrat po příčině neshod mezi vzniklými genovými stromy. Osekvenováno by mělo být okolo tisíce jaderných lokusů několika desítek druhů rodu *Curcuma* L. pomocí NGS přístupu, konkrétně Hyb-Seq metody (Weitemier et al., 2014). Vzorky budou vybírány tak, aby pokryly rodovou variabilitu včetně druhů, které byly navrženy v práci Záveská et al. (manuscript) jako meziskupinoví hybridi.

Hyb-Seq je metoda kombinující cílové obohacení a mělké sekvenování, čímž jsou získána data jak z nízkokopiových jaderných markerů, tak vysokokopiových lokusů. K návrhu obohacovacích sond se využívají genomová a transkriptomová data, z nichž jsou vybrány unikátní přepisované oblasti. Po hybridizaci s próbami v roztoku a Illumina sekvenování jsou ready namapovány na pseudoreferenci – sekvenci obohacovaných exonů oddělených 200 N pozic. Exony jsou zřetězeny do genů a sestavena sekvence plastomu a rDNA. Následuje samotné odvození genových stromů a odvození celkové fylogeneze dané skupiny (Weitemier et al., 2014).

4618 exonů z 1180 genů druhu *Curcuma longa* L. bylo vybráno pro přípravu hybridizačních sond. Asi pět druhů *Curcuma* L. již bylo osekvenováno, data vykazují pouze 5 % chybějících údajů a v pseudoreferenci byl zvýšen počet N pozic na 400 z důvodu přesahu do okolních oblastí při použití pseudoreference obsahující pouze 200 N pozic (Fér T. unpubl. data).

6 Závěr

Next-generation sekvenování přineslo velký rozvoj v analýze genomů a transkriptomů. Umožnilo rychle a jednoduše osekvenovat obrovské množství vzorků za mnohem nižší cenu, než je běžné u Sangerovy metody. Celogenomové sekvenování je však stále nedostupné pro mnohé studie založené na větším počtu jedinců, a proto se vyvinulo mnoho metod, umožňujících snížit genomovou komplexitu a sloučit jednotlivé vzorky do jedné sekvenační reakce tak, aby při následné analýze dat bylo možno je opět rozlišit.

Každá metoda na snížení genomové komplexity má několik výhod, ale také nevýhod. Amplikonové sekvenování je finančně náročné pro velké studie a trpí problémy spojenými s metodou PCR, na níž je založeno. Restrikční metody vhodné pro nemodelové organismy nejsou využitelné v hlubších fylogenetických analýzách. Pro použití cíleného obohacení je nutné znát alespoň částečnou referenční sekvenci, lze jím však jednoduše cílit na informativní lokusy. Pomocí sekvenování transkriptomů lze jednoduše zacílit na přepisované části genomu a stejně jako restrikční metody nevyžaduje předešlou znalost cílových sekvencí. V opozici k tomu stojí však složitější práce s RNA oproti DNA, rozdílná transkripce mezi jednotlivými orgány, vývojovými stádii a náročná assembly kvůli alternativnímu sestřihu. Mělké sekvenování patří k jednoduchým metodám, ale fylogenetické analýzy znesnadňuje častý rozpor mezi stromy získanými z různých genomů a problémy s využitím vysokokopiových částí DNA u polyploidních druhů.

Fylogenetika polyploidních druhů oproti odvozování evolučních historií druhů diploidních skýtá několik úskalí. Kvůli uniparentální dědičnosti cytoplazmatických genomů a tíhnutí k homogenizaci homeologních lokusů rDNA je značně omezena využitelnost v minulosti asi nejčastěji využívaných markerů – chloroplastové DNA a ITS. Vývoj markerů z nízkokopiové jaderné frakce DNA byl při využití klasických molekulárních technik značně zdlouhavý a finančně náročný. Tuto nevýhodu překonává next-generation sekvenování, které umožňuje osekvenovat i tisíce různých lokusů několika až několika set jedinců, čímž řeší i druhý častý dřívější problém – nedostatek dat (různých nezávislých lokusů) pro fylogenetické analýzy. Avšak práce s obrovským NGS datasetem je stále náročná jak na výpočetní zdroje, tak na vývoj nových modelů a hypotéz, které poslouží jako podklad pro následné analýzy.

Seznam literatury

Adams, K.L., and Wendel, J.F. (2005). Polyploidy and genome evolution in plants. Current Opinion in Plant Biology *8*, 135–141.

Adessi, C., Matton, G., Ayala, G., et al. (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res. *28*, E87.

Altshuler, D., Pollara, V.J., Cowles, C.R., et al. (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature *407*, 513–516.

Álvarez, I., and Wendel, J.F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. Molecular Phylogenetics and Evolution *29*, 417–434.

Arnold, B., Kim, S.-T., and Bomblies, K. (2015). Single Geographic Origin of a Widespread Autotetraploid Arabidopsis arenosa Lineage Followed by Interploidy Admixture. Molecular Biology and Evolution *32*, 1382–1395.

Baird, N.A., Etter, P.D., Atwood, T.S., et al. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE *3*, e3376.

Batley, J. (2003). Mining for Single Nucleotide Polymorphisms and Insertions/Deletions in Maize Expressed Sequence Tag Data. PLANT PHYSIOLOGY *132*, 84–91.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Bi, K., Vanderpool, D., Singhal, S., et al. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics *13*, 403.

Binladen, J., Gilbert, M.T.P., Bollback, J.P., et al. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS ONE *2*, e197.

Bock, D.G., Kane, N.C., Ebert, D.P., and Rieseberg, L.H. (2014). Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. New Phytologist *201*, 1021–1030.

Bombarely, A., Edwards, K.D., Sanchez-Tamburrino, J., and Mueller, L.A. (2012). Deciphering the complex leaf transcriptome of the allotetraploid species Nicotiana tabacum: a phylogenomic perspective. BMC Genomics *13*, 406.

Bombarely, A., Coate, J.E., and Doyle, J.J. (2014). Mining transcriptomic data to study the origins and evolution of a plant allopolyploid complex. PeerJ *2*, e391.

Briggs, A.W., Good, J.M., Green, R.E., et al. (2009). Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. Science *325*, 318–321.

Bryant, D., and Moulton, V. (2004). Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. Molecular Biology and Evolution *21*, 255–265.

Brysting A. K., Mathiesen C., and Marcussen T. (2011). Challenges in polyploid phylogenetic reconstruction: A case story from the arctic-alpine Cerastium alpinum complex. Taxon *60*, 333–347.

Bybee, S.M., Bracken-Grissom, H., Haynes, B.D., et al. (2011). Targeted Amplicon Sequencing (TAS): A Scalable Next-Gen Approach to Multilocus, Multitaxa Phylogenetics. Genome Biology and Evolution *3*, 1312–1323.

Camin, J.H., and Sokal, R.R. (1965). A Method for Deducing Branching Sequences in Phylogeny. Evolution *19*, 311.

Cannon, S.B., McKain, M.R., Harkess, A., et al. (2015). Multiple Polyploidy Events in the Early Radiation of Nodulating and Nonnodulating Legumes. Molecular Biology and Evolution *32*, 193–210.

Cavanagh, C.R., Chao, S., Wang, S., et al. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proceedings of the National Academy of Sciences *110*, 8057–8062.

Chamberlain, J.S., Gibbs, R.A., Rainer, J.E., et al. (1988). Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. Nucleic Acids Research *16*, 11141–11156.

Chan, Y.C., Roos, C., Inoue-Murayama, M., et al. (2010). Mitochondrial Genome Sequences Effectively Reveal the Phylogeny of Hylobates Gibbons. PLoS ONE *5*, e14419.

Clevenger, J., Chavarro, C., Pearl, S.A., et al. (2015). Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. Mol Plant.

Coyne, K.J., Burkholder, J.M., Feldman, R.A., et al. (2004). Modified serial analysis of gene expression method for construction of gene expression profiles of microbial eukaryotic species. Appl. Environ. Microbiol. *70*, 5298–5304.

Cronn, R., Cedroni, M., Haselkorn, T., et al. (2002). PCR-mediated recombination in amplification products derived from polyploid cotton. Theor. Appl. Genet. *104*, 482–489.

Cronn, R., Knaus, B.J., Liston, A., et al. (2012). Targeted enrichment strategies for next-generation plant biology. American Journal of Botany *99*, 291–311.

Cronn, R.C., Small, R.L., and Wendel, J.F. (1999). Duplicated genes evolve independently after polyploid formation in cotton. Proc. Natl. Acad. Sci. U.S.A. *96*, 14406–14411.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., et al. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics *12*, 499–510.

Degnan, J.H., and Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution *24*, 332–340.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends in Genetics *30*, 418–426.

Dressman, D., Yan, H., Traverso, G., et al. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc. Natl. Acad. Sci. U.S.A. *100*, 8817–8822.

Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution *29*, 1969–1973.

Dufresne, F., Stift, M., Vergilino, R., and Mable, B.K. (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. Molecular Ecology *23*, 40–69.

Edwards, A.W.F. (2009). Statistical Methods for Evolutionary Trees. Genetics *183*, 5–12.

* Edwards, A.W.F., Cavalli-Sforza, L.L., (1964). Reconstruction of evolutionary trees, pp. 67–76 in: V.H. Heywood and J. McNeill, eds., Phenetic and Phylogenetic Classification. London: Systematics Association.

Egan, A.N., Schlueter, J., and Spooner, D.M. (2012). Applications of next-generation sequencing in plant biology. American Journal of Botany *99*, 175–185.

Eid, J., Fehr, A., Gray, J., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138.

Elshire, R.J., Glaubitz, J.C., Sun, Q., et al. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE *6*, e19379.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., et al. (2012). Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. Systematic Biology *61*, 717–726.

Fortune, P.M., Pourtau, N., Viron, N., and Ainouche, M.L. (2008). Molecular phylogeny and reticulate origins of the polyploid Bromus species from section Genea (Poaceae). American Journal of Botany *95*, 454–464.

Fu, Y., Springer, N.M., Gerhardt, D.J., et al. (2010). Repeat subtraction-mediated sequence capture from a complex genome: Sequence capture in maize. The Plant Journal *62*, 898–909.

Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. Molecular Ecology Resources *11*, 759–769.

Gnirke, A., Melnikov, A., Maguire, J., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology *27*, 182–189.

Govindarajulu, R., Parks, M., Tennessen, J.A., et al. (2015). Comparison of nuclear, plastid, and mitochondrial phylogenies and the origin of wild octoploid strawberry species. American Journal of Botany *102*, 544–554.

Griffin, P.C., Robin, C., and Hoffmann, A.A. (2011). A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to Poa grasses. BMC Biology *9*, 19.

Grover, C.E., Zhu, X., Grupp, K.K., et al. (2015). Molecular confirmation of species status for the allopolyploid cotton species, Gossypium ekmanianum Wittmack. Genetic Resources and Crop Evolution *62*, 103–114.

Hand, M.L., Cogan, N.O.I., and Forster, J.W. (2012). Genome-wide SNP identification in multiple morphotypes of allohexaploid tall fescue (Festuca arundinacea Schreb). BMC Genomics *13*, 219.

Hovav, R., Udall, J.A., Chaudhary, B., et al. (2008). Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. Proc. Natl. Acad. Sci. U.S.A. *105*, 6191–6195.

Huson, D.H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. *23*, 254–267.

Hutchison, C.A. (2007). DNA sequencing: bench to bedside and beyond. Nucleic Acids Research *35*, 6227–6237.

Jarne, P., and Lagoda, P.J. (1996). Microsatellites, from molecules to populations and back. Trends Ecol. Evol. (Amst.) *11*, 424–429.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature *473*, 97–100.

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., et al. (2012). A genome triplication associated with early diversification of the core eudicots. Genome Biology *13*, R3.

Kumar, S., Sood, A., Wegener, J., et al. (2005). TERMINAL PHOSPHATE LABELED NUCLEOTIDES: SYNTHESIS, APPLICATIONS, AND LINKER EFFECT ON

INCORPORATION BY DNA POLYMERASES. Nucleosides, Nucleotides and Nucleic Acids *24*, 401–408.

Lam, H.-M., Xu, X., Liu, X., et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nature Genetics *42*, 1053–1059.

Lemmon, E.M., and Lemmon, A.R. (2013). High-Throughput Genomic Data in Systematics and Phylogenetics. Annual Review of Ecology, Evolution, and Systematics *44*, 99–121.

Lennon, N.J., Lintner, R.E., Anderson, S., et al. (2010). A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. Genome Biology *11*, R15.

Li, X., Acharya, A., Farmer, A.D., et al. (2012). Prevalence of single nucleotide polymorphism among 27 diverse alfalfa genotypes as assessed by transcriptome sequencing. BMC Genomics *13*, 568.

Linder, C.R., and Rieseberg, L.H. (2004). Reconstructing patterns of reticulate evolution in plants. Am. J. Bot. *91*, 1700–1708.

Lott, M., Spillner, A., Huber, K.T., et al. (2009). Inferring polyploid phylogenies from multiply-labeled gene trees. BMC Evolutionary Biology *9*, 216.

Lu, F., Lipka, A.E., Glaubitz, J., et al. (2013). Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. PLoS Genetics *9*, e1003215.

Maddison, W.P. (1997). Gene Trees in Species Trees. Systematic Biology *46*, 523.

Mandáková, T., Joly, S., Krzywinski, M., et al. (2010). Fast Diploidization in Close Mesopolyploid Relatives of Arabidopsis. THE PLANT CELL ONLINE *22*, 2277–2290.

Margulies, M., Egholm, M., Altman, et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature.

Maricic, T., Whitten, M., and Pääbo, S. (2010). Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. PLoS ONE *5*, e14004.

Marioni, J.C., Mason, C.E., Mane, S.M., et al. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research *18*, 1509–1517.

Markoulatos, P., Siafakas, N., and Moncany, M. (2002). Multiplex polymerase chain reaction: a practical approach. J. Clin. Lab. Anal. *16*, 47–51.

Maxam, A.M., and Gilbert, W. (1977). A new method for sequencing DNA. Proc. Natl. Acad. Sci. U.S.A. *74*, 560–564.

McCormack, J.E., Hird, S.M., Zellmer, A.J., et al. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution *66*, 526–538.

McKain, M.R., Wickett, N., Zhang, Y., et al. (2012). Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). American Journal of Botany *99*, 397–406.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nature Reviews Genetics *11*, 31–46.

Meyer, M., Stenzel, U., and Hofreiter, M. (2008). Parallel tagged sequencing on the 454 platform. Nature Protocols *3*, 267–278.

Morin, R., Bainbridge, M., Fejes, A., et al. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques *45*, 81–94.

Mutter, G.L., and Boynton, K.A. (1995). PCR bias in amplification of androgen receptor alleles, a trinucleotide repeat marker used in clonality studies. Nucleic Acids Res. *23*, 1411–1418.

Njuguna, W., Liston, A., Cronn, R., and Bassil, N. (2010). Multiplexed Fragaria Chloroplast Genome Sequencing. Acta Horticulturae 315–321.

O'Hanlon, P.C., and Peakall, R. (2000). A simple method for the detection of size homoplasy among amplified fragment length polymorphism fragments. Mol. Ecol. *9*, 815–816.

Okou, D.T., Steinberg, K.M., Middle, C., et al. (2007). Microarray-based genomic selection for high-throughput resequencing. Nature Methods *4*, 907–909.

O'Neill, E.M., Schwartz, R., Bullock, C.T., et al. (2013). Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. Molecular Ecology *22*, 111–129.

van Orsouw, N.J., Hogers, R.C.J., Janssen, A., et al. (2007). Complexity Reduction of Polymorphic Sequences (CRoPS^{TM}): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. PLoS ONE *2*, e1172.

Parks, M., Cronn, R., and Liston, A. (2009). Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. *7*, 84.

Peterson, B.K., Weber, J.N., Kay, E.H., et al. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE *7*, e37135.

Pfeil, B.E., Brubaker, C.L., Craven, L.A., and Crisp, M.D. (2004). Paralogy and orthology in the MALVACEAE rpb2 gene family: investigation of gene duplication in hibiscus. Mol. Biol. Evol. *21*, 1428–1437.

Pickrell, J.K., and Pritchard, J.K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genetics *8*, e1002967.

Porreca, G.J., Zhang, K., Li, J.B., et al. (2007). Multiplex amplification of large sets of human exons. Nat. Methods *4*, 931–936.

Portnoy, V., Diber, A., Pollock, S., et al. (2011). Use of Non-Normalized, Non-Amplified cDNA for 454-Based RNA Sequencing of Fleshy Melon Fruit. The Plant Genome Journal *4*, 36.

Rabinowicz, P.D., Schutz, K., Dedhia, N., et al. (1999). Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat. Genet. *23*, 305–308.

Rannala, B., and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. *43*, 304–311.

Rastogi, S., and Liberles, D.A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol. Biol. *5*, 28.

Rauwolf, U., Golczyk, H., Greiner, S., and Herrmann, R.G. (2010). Variable amounts of DNA related to the size of chloroplasts III. Biochemical determinations of DNA amounts per organelle. Mol. Genet. Genomics *283*, 35–47.

Richardson, B.A., Page, J.T., Bajgain, P., et al. (2012). Deep sequencing of amplicons reveals widespread intraspecific hybridization and multiple origins of polyploidy in big sagebrush (Artemisia tridentata; Asteraceae). American Journal of Botany *99*, 1962–1975.

Roberts, J.D., Preston, B.D., Johnston, L.A., et al. (1989). Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis in vitro. Mol. Cell. Biol. *9*, 469–476.

Rogers, S.O., and Bendich, A.J. (1987). Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. Plant Molecular Biology *9*, 509–520.

Ronaghi, M., Karamohamed, S., Pettersson, B., et al. (1996). Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem. *242*, 84–89.

Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572–1574.

Rothberg, J.M., Hinz, W., Rearick, T.M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature *475*, 348–352.

Rousseau-Gueutin, M., Gaston, A., Aïnouche, A., et al. (2009). Tracking the evolutionary history of polyploidy in Fragaria L. (strawberry): New insights from phylogenetic analyses of low-copy nuclear genes. Molecular Phylogenetics and Evolution *51*, 515–530.

Roychowdhury, T., Vishnoi, A., and Bhattacharya, A. (2013). Next-Generation Anchor Based Phylogeny (NexABP): Constructing phylogeny from Next-generation sequencing data. Scientific Reports *3*.

Rubin, B.E.R., Ree, R.H., and Moreau, C.S. (2012). Inferring Phylogenies from RAD Sequence Data. PLoS ONE *7*, e33394.

Sachidanandam, R., Weissman, D., Schmidt, S.C., et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928–933.

Saintenac, C., Jiang, D., and Akhunov, E.D. (2011). Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biology *12*, R88.

Sang, T. (2002). Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit. Rev. Biochem. Mol. Biol. *37*, 121–147.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. *74*, 5463–5467.

Schneeberger, K., Ossowski, S., Ott, F., et al. (2011). Reference-guided assembly of four diverse Arabidopsis thaliana genomes. Proceedings of the National Academy of Sciences *108*, 10249–10254.

* Schumann, K. (1904). Zingiberaceae. In Das Pflanzenreich IV, Engler, A., ed. (Leipzig: Engelmann), pp. 1–458.

Severin, A.J., Woody, J.L., Bolon, Y.-T., et al. (2010). RNA-Seq Atlas of Glycine max: A guide to the soybean transcriptome. BMC Plant Biology *10*, 160.

Škorničková, J (2007). Taxonomic Studies in Indian Curcuma L. PhD. Thesis. Charles University, Prague.

Small, R.L., Cronn, R.C., and Wendel, J.F. (2004). L. A. S. JOHNSON REVIEW No. 2. Use of nuclear genes for phylogeny reconstruction in plants. Australian Systematic Botany *17*, 145.

Smedmark, J.E.E., Eriksson, T., Evans, R.C., and Campbell, C.S. (2003). Ancient allopolyploid speciation in Geinae (Rosaceae): evidence from nuclear granule-bound starch synthase (GBSSI) gene sequences. Syst. Biol. *52*, 374–385.

Sousa, F. de, Bertrand, Y.J.K., Nylinder, S., et al. (2014). Phylogenetic Properties of 50 Nuclear Loci in Medicago (Leguminosae) Generated Using Multiplexed Sequence Capture and Next-Generation Sequencing. PLoS ONE *9*, e109704.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Straub, S.C., Fishbein, M., Livshultz, T., et al. (2011). Building a model: developing genomic resources for common milkweed (Asclepias syriaca) with low coverage genome sequencing. BMC Genomics *12*, 211.

Straub, S.C.K., Parks, M., Weitemier, K., et al. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. American Journal of Botany *99*, 349–364.

Stull, G.W., Moore, M.J., Mandala, V.S., et al. (2013). A Targeted Enrichment Strategy for Massively Parallel Sequencing of Angiosperm Plastid Genomes. Applications in Plant Sciences *1*, 1200497.

Syring, J., Farrell, K., Businský, R., et al. (2007). Widespread Genealogical Nonmonophyly in Species of Pinus Subgenus Strobus. Systematic Biology *56*, 163–181.

Swofford, D. L. (2002). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Van Tassell, C.P., Smith, T.P.L., Matukumalli, L.K., et al. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods *5*, 247–252.

Tennessen, J.A., Govindarajulu, R., Ashman, T.L., and Liston, A. (2014). Evolutionary Origins and Dynamics of Octoploid Strawberry Subgenomes Revealed by Dense Targeted Capture Linkage Maps. Genome Biology and Evolution *6*, 3295–3313.

Tewhey, R., Warner, J.B., Nakano, M., et al. (2009a). Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nature Biotechnology *27*, 1025–1031.

Tewhey, R., Nakano, M., Wang, X., et al. (2009b). Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biology *10*, R116.

Tonnabel, J., Olivieri, I., Mignot, A., et al. (2014). Developing nuclear DNA phylogenetic markers in the angiosperm genus Leucadendron (Proteaceae): A next-generation sequencing transcriptomic approach. Molecular Phylogenetics and Evolution *70*, 37–46.

Turner, E.H., Lee, C., Ng, S.B., et al. (2009). Massively parallel exon capture and library-free resequencing across 16 genomes. Nature Methods *6*, 315–316.

Uitdewilligen, J.G.A.M.L., Wolters, A.M.A., D'hoop, B.B., et al. (2013). A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. PLoS ONE *8*, e62355.

Utturkar, S.M., Klingeman, D.M., Land, M.L., et al. (2014). Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics *30*, 2709–2716.

Valouev, A., Ichikawa, J., Tonthat, T., et al. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Research *18*, 1051–1063.

Vekemans, X., Beauwens, T., Lemaire, M., and Roldán-Ruiz, I. (2002). Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Mol. Ecol. *11*, 139–151.

Vos, P., Hogers, R., Bleeker, M., et al. (1995). AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. *23*, 4407–4414.

Wang, X.Q., Zhao, L., Eaton, D.A.R., et al. (2013). Identification of SNP markers for inferring phylogeny in temperate bamboos (Poaceae: Bambusoideae) using RAD sequencing. Molecular Ecology Resources *13*, 938–945.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics *10*, 57–63.

Weitemier, K., Straub, S.C.K., Cronn, R.C., et al. (2014). Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant Phylogenomics. Applications in Plant Sciences *2*, 1400042.

Wendel, J.F., Schnabel, A., and Seelanan, T. (1995). Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (Gossypium). Proc. Natl. Acad. Sci. U.S.A. *92*, 280–284.

Yi, H., and Jin, L. (2013). Co-phylog: an assembly-free phylogenomic approach for closely related organisms. Nucleic Acids Research *41*, e75–e75.

Záveská, E., Fér, T., Šída, O., et al. (2011). Genetic diversity patterns in Curcuma reflect differences in genome size: GENETIC DIVERSITY PATTERNS IN CURCUMA. Botanical Journal of the Linnean Society *165*, 388–401.

Záveská, E. (2014). Phylogenetic Studies in the Polyploid Genus Curcuma L. Ph.D. Thesis, Charles University, Prague.

Záveská, E., Fér, T., Šída, O., et al. (manuscript). Hybridization among distantly related species: examples from the polyploid genus Curcuma (Zingiberaceae).

Záveská, E., Fér, T., Šída, O., et al. (2012). Phylogeny of Curcuma (Zingiberaceae) based on plastid and nuclear sequences: proposal of the new subgenus Ecomata. Taxon *61*, 747–763.

Zeng, X.-C., and Wang, S.-X. (2002). Evidence that BmTXK beta-BmKCT cDNA from Chinese scorpion Buthus martensii Karsch is an artifact generated in the reverse transcription process. FEBS Lett. *520*, 183–184; author reply 185.

Zhulidov, P.A., Bogdanova, E.A., Shcheglov, A.S., et al. (2004). Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Res. *32*, e37.

Internetové zdroje

Glenn, Travis (2014). 2014 NGS Field Guide: Overview [online], [cit. 2015-4-03]. Dostupné z: http://www.molecularecologist.com/next-gen-fieldguide-2014/

Illumina [online], [cit. 2015-3-29]. Dostupné z: http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Life Technologies[1] [online], [cit. 2015-3-31]. Dostupné z: http://www.lifetechnologies.com/cz/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html

Life Technologies[2] [online], [cit. 2015-3-29]. Dostupné z: http://www.lifetechnologies.com/cz/en/home/life-science/sequencing/next-generation-sequencing/solid-next-generation-sequencing/solid-next-generation-sequencing-systems-reagents-accessories/solid-next-generation-sequencing-chemistry.html

Pacific Biosciences [online], [cit. 2015-4-02]. Dostupné z: http://www.pacificbiosciences.com/products/smrt-technology/

Roche [online], [cit. 2015-3-28]. Dostupné z: http://www.454.com/products/technology.asp

The 1000 Plants Project [online], [cit. 2015-5-30]. Dostupné z: http://www.onekp.com